PCT WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 6 :
C12N 15/52, C12P 17/18, 17/10, C12N 9/00, 1/21, 15/70, C12Q 1/68

A1

(11) International Publication Number: **WO 98/07868**

(43) International Publication Date: 26 February 1998 (26.02.98)

(21) International Application Number: PCT/EP97/04495

(22) International Filing Date: 18 August 1997 (18.08.97)

(30) Priority Data:
96810551.0 20 August 1996 (20.08.96) EP
(34) Countries for which the regional or international application was filed: DE et al.

(71) Applicant *(for all designated States except US)*: NOVARTIS AG [CH/CH]; Schwarzwaldallee 215, CH-4058 Basel (CH).

(72) Inventors; and
(75) Inventors/Applicants *(for US only)*: SCHUPP, Thomas [CH/CH]; Fritschmattweg 5, CH-4313 Möhlin (CH). TOUPET, Christiane [FR/FR]; 12/5, rue de l'Ours, F-68200 Mulhouse (FR). ENGEL, Nathalie [FR/FR]; 29, rue de la Doller, F-68260 Kingersheim (FR).

(74) Agent: ROTH, Bernhard, M.; Novartis AG, Patent- und Markenabteilung, Lichtstrasse 35, CH-4002 Basel (CH).

(81) Designated States: AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, CA, CH, CN, CU, CZ, DE, DK, EE, ES, FI, GB, GE, GH, HU, IL, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MD, MG, MK, MN, MW, MX, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, UA, UG, US, UZ, VN, YU, ZW, ARIPO patent (GH, KE, LS, MW, SD, SZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, ML, MR, NE, SN, TD, TG).

Published
With international search report.
Before the expiration of the time limit for amending the claims and to be republished in the event of the receipt of amendments.

(54) Title: RIFAMYCIN BIOSYNTHESIS GENE CLUSTER

(57) Abstract

The present invention primarily relates to a DNA fragment which is obtainable from the gene cluster responsible for rifamycin biosynthesis within the genome of *Amycolatopsis mediterranei*, and comprises at least one gene or a part of a gene which codes for a polypeptide which is directly or indirectly involved in the biosynthesis of rifamycin, and to a method for preparing said DNA fragment. The present invention furthermore relates to recombinant DNA molecules which comprise one of the DNA fragments according to the invention, and to the plasmids and vectors derived therefrom. Host organisms transformed with said plasmid or vector DNA are likewise embraced.

FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AL	Albania	ES	Spain	LS	Lesotho	SI	Slovenia
AM	Armenia	FI	Finland	LT	Lithuania	SK	Slovakia
AT	Austria	FR	France	LU	Luxembourg	SN	Senegal
AU	Australia	GA	Gabon	LV	Latvia	SZ	Swaziland
AZ	Azerbaijan	GB	United Kingdom	MC	Monaco	TD	Chad
BA	Bosnia and Herzegovina	GE	Georgia	MD	Republic of Moldova	TG	Togo
BB	Barbados	GH	Ghana	MG	Madagascar	TJ	Tajikistan
BE	Belgium	GN	Guinea	MK	The former Yugoslav Republic of Macedonia	TM	Turkmenistan
BF	Burkina Faso	GR	Greece			TR	Turkey
BG	Bulgaria	HU	Hungary	ML	Mali	TT	Trinidad and Tobago
BJ	Benin	IE	Ireland	MN	Mongolia	UA	Ukraine
BR	Brazil	IL	Israel	MR	Mauritania	UG	Uganda
BY	Belarus	IS	Iceland	MW	Malawi	US	United States of America
CA	Canada	IT	Italy	MX	Mexico	UZ	Uzbekistan
CF	Central African Republic	JP	Japan	NE	Niger	VN	Viet Nam
CG	Congo	KE	Kenya	NL	Netherlands	YU	Yugoslavia
CH	Switzerland	KG	Kyrgyzstan	NO	Norway	ZW	Zimbabwe
CI	Côte d'Ivoire	KP	Democratic People's Republic of Korea	NZ	New Zealand		
CM	Cameroon			PL	Poland		
CN	China	KR	Republic of Korea	PT	Portugal		
CU	Cuba	KZ	Kazakstan	RO	Romania		
CZ	Czech Republic	LC	Saint Lucia	RU	Russian Federation		
DE	Germany	LI	Liechtenstein	SD	Sudan		
DK	Denmark	LK	Sri Lanka	SE	Sweden		
EE	Estonia	LR	Liberia	SG	Singapore		

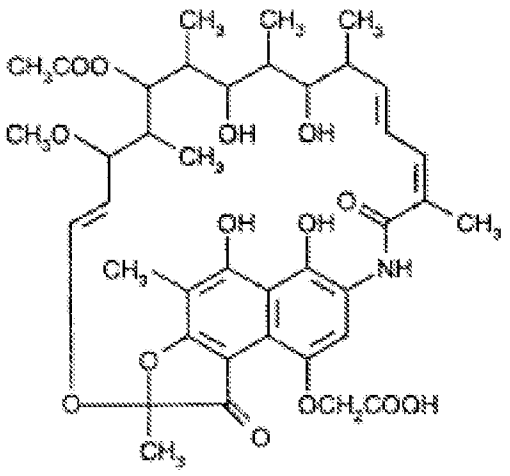
Rifamycin biosynthesis gene cluster

Rifamycins form an important group of macrocyclic antibiotics (Wehrli, Topics in Current Chemistry (1971), 72, 21-49). They consist of a naphthoquinone chromophore which is spanned by a long aliphatic bridge. Rifamycins belong to the class of ansamycin antibiotics which are produced by several Gram-positive soil bacteria of the actinomycetes group and a few plants.

Ansamycins are characterized by a flat aromatic nucleus spanned by a long aliphatic bridge joining opposite positions of the nucleus. Two different groups of ansamycins can be distinguished by the structure of the aromatic nucleus. One group has a naphthoquinoid chromophore, with the typical representatives being rifamycin, streptovaricin, tolypomycin and naphthomycin. The second group, which has a benzoquinoid chromophore, is characterized by geldanamycin, maytansines and ansamitocines (Ghisalba et al., Biotechnology of Industrial Antibiotics Vandamme E. J. Ed., Decker Inc. New York, (1984) 281-327). In contrast to antibiotics of the macrolide type, the ansamycins contain in the aliphatic ring system not a lactone linkage but an amide linkage which forms the connection to the chromophore.

The discovery of the rifamycins produced by the microorganism *Streptomyces mediterranei* (as the organism was called at that time, see below) was described for the first time in 1959 (Sensi et al., Farmaco Ed. Sci. (1959) 14, 146-147). Extraction with ethyl acetate of the acidified cultures of *Streptomyces mediterranei* resulted in isolation of a mixture of antibiotically active components, the rifamycins A, B, C, D and E. Rifamycin B, the most stable component, was separated from the other components and isolated on the basis of its strongly acidic properties and ease of salt formation.

Rifamycin B has the structure of the formula (1)



(1)

Rifamycin B is the main component of the fermentation when barbiturate is added to the fermentation medium and/or improved producer mutants of *Streptomyces mediterranei* are used.

The rifamycin producer strain was originally classified as *Streptomyces mediterranei* (Sensi et al., Farmaco Ed. Sci. (1959) **14**, 146-147). Analysis of the cell wall of *Streptomyces mediterranei* by Thiemann et al. later revealed that this strain has a cell wall typical of *Nocardia*, and the strain was reclassified as *Nocardia mediterranei* (Thieman et al. Arch. Microbiol. (1969), **67** 147-151). *Nocardia mediterranei* has been reclassified again on the basis of more recent accurate morphological and biochemical criteria. Based on the exact composition of the cell wall, the absence of mycolic acid and the insensitivity to *Nocardia* and *Rhodococcus* phages, the strain has been assigned to the new genus *Amycolatopsis* as *Amycolatopsis mediterranei* (Lechevalier et al., Int. J. Syst. Bacteriol. (1986), 36, 29).

Rifamycins have a strong antibiotic activity mainly against Gram-positive bacteria such as mycobacteria, neisserias and staphylococci. The bactericidal effect of rifamycins derives from specific inhibition of the bacterial DNA-dependent RNA polymerase, which interrupts RNA biosynthesis (Wehrli and Staehelin, Bacteriol. Rev. (1971), **35**, 290-309). The semisynthetic rifamycin B derivative rifampin (rifampicin) is widely used clinically as antibiotic against the agent causing tuberculosis, *Mycobacterium tuberculosis*.

The naphthoquinoid ansamycins of the streptovaricin and tolypomycin group show, like rifamycin, an antibacterial effect by inhibiting bacterial RNA polymerase. By contrast, naphthomycin has an antibacterial effect without inhibiting bacterial RNA polymerase. The

benzoquinoid ansamycins show no inhibition of bacterial RNA polymerase, and they therefore have only relatively weak antibacterial activity, if any. On the other hand, some representatives of this class of substances have an effect on eukaryotic cells. Thus, antifungal, antiprotozoal and antitumour properties have been described for geldanamycin. On the other hand, antimitotic (antitubilin), antileukaemic and antitumour properties are ascribed to the maytansines. Some rifamycins also show antitumour and antiviral activity, but only at high concentrations. This biological effect thus appears to be nonspecific.

Despite the great structural variety of the ansamycins, their biosynthesis appears to take place by a metabolic pathway which contains many common elements (Ghisalba et al. Biotechnology of Industrial Antibiotics Vandamme E. J. Ed., Decker Inc. New York, (1984) 281-327). The aromatic nucleus for all ansamycins is probably built up starting from 3-amino-5-hydroxybenzoic acid. Starting from this molecule, which is presumably activated as coenzyme A, the entire aliphatic bridge is synthesized by a multifunctional polyketide synthase. The length of the bridge and the processing of the keto groups, which are initially formed by the condensation steps, are controlled by the polyketide synthase. To build up the complete aliphatic bridge for rifamycins, 10 condensation steps, 2 with acetate and 8 with propionate as building blocks, are necessary. The sequence of these individual condensation steps is likewise determined by the polyketide synthase. Structural comparisons and studies with incorporation of radioactive acetate and propionate have shown that the sequence of acetate and propionate incorporation for the various ansamycins takes place in accordance with a scheme which appears to be identical or very similar in the first condensation steps. Thus, from a common synthesis scheme of the ansamycin polyketide synthases (the rifamycin synthesis scheme), the syntheses of the various ansamycins sooner or later branch off, in accordance with their structural difference from the rifamycin structure, into side branches of the synthesis (Ghisalba et al., Biotechnology of Industrial Antibiotics Vandamme E. J. Ed., Decker Inc. New York, (1984) 281-327).

Because of the great structural variety of the rifamycins and their specific and interesting biological effect, there is great interest in understanding the genetic basis of their synthesis in order to create the possibility of specifically influencing it. This is particularly desirable because, as explained above, there is much in common between the synthesis of rifamycins and that of other ansamycins. This similarity in the biosynthesis, which probably derives from a common evolutionary origin of this metabolic pathway, naturally has a genetic basis.

The genetic basis of secondary metabolite biosynthesis essentially exists in the genes which code for the individual biosynthetic enzymes, and in the regulatory elements which control the expression of the biosynthesis genes. The secondary metabolite synthesis genes of actinomycetes have hitherto been found as clusters of adjacent genes in all the systems investigated. The size of such antibiotic gene clusters extends from about 10 kilobases (kb) up to more than 100 kb. The clusters often contain specific regulator genes and genes for resistance of the producer organism to its own antibiotic (Chater, Ciba Found. Symp. (1992), 171, 144-162).

The invention described herein has now succeeded, by identifying and cloning genes of rifamycin biosynthesis, in creating the genetic basis for synthesizing by genetic methods rifamycin analogues or novel ansamycins which combine structural elements from rifamycin with other ansamycins. This also creates the basis for preparing novel collections of substances based on the rifamycin biosynthesis gene cluster by combinatorial biosynthesis.

It was possible in a first step to identify and clone a DNA fragment from the genome of *A. mediterranei*, which shows homology with known polyketides synthase genes. After obtaining the sequence information from this DNA fragment which confirmed a typical sequence for polyketide synthases it was possible to screen a cosmid library of *A. mediterranei* with specific DNA probes derived from this fragment in a screening program for further DNA fragments which are involved in the rifamycin gene cluster. As a result, the complete rifamycin polyketide synthase gene cluster was identified and subjected to sequence determination (see SEQ ID NO 3). The gene cluster comprises six open reading frames, which are referred to hereinafter as ORF A, B, C, D, E and F and which code for the proteins and polypeptides depicted in SEQ ID NOS 4 to 9.

The gene cluster isolated and characterized in this way represents the basis, for example, for targeted optimization of the production of rifamycin, ansamycins or analogues thereof. Examples of techniques and possible areas of application available in this connection are as follows:

- Overexpression of individual genes in producer strains with plasmid vectors or by incorporation into the chromosome.
- Study of the expression and transcriptional regulation of the gene cluster during fermentation with various producer strains and optimization thereof through physiological parameters and appropriate fermentation conditions.

- Identification of regulatory genes and of the DNA binding sites of the corresponding regulatory proteins in the gene cluster. Characterization of the effect of these regulatory elements on the production of rifamycins or ansamycins; and influencing them by specific mutation in these genes or the DNA binding sites.
- Duplication of the complete gene cluster or parts thereof in producer strains.

Besides these applications of the gene cluster to improve production by fermentation as described above, it can likewise be employed for the biosynthetic preparation of novel rifamycin analogues or novel ansamycins or ansamycin-like compounds in which the aliphatic bridge is connected at only one end to the aromatic nucleus. The following possibilities come into consideration here, for example:

- Inactivation of individual steps in the biosynthesis, for example by gene disruption.
- Mutation of individual steps in the biosynthesis, for example by gene replacement.
- Use of the cluster or fragments thereof as DNA probe in order to isolate other natural microorganisms which produce metabolites similar to rifamycin or ansamycins.
- Exchange of individual elements in this gene cluster by those from other gene clusters.
- Use of modified polyketide synthases for setting up libraries of various rifamycin analogues or ansamycins, which are then tested for their activity (Jackie & Khosla, Chemistry & Biology, (1995), 2, 355-362).
- Construction of mutated actinomycetes strains from which the natural rifamycin or ansamycin biosynthesis gene cluster in the chromosome has been partly or completely deleted, and can thus be used for expressing genetically modified gene clusters.
- Exchange of individual elements within the gene cluster.

Detailed description of the invention

The invention relates to a DNA fragment from the genome of *Amycolatopsis mediterranei*, which comprises a DNA region which is involved directly or indirectly in the gene cluster responsible for rifamycin synthesis; and the adjacent DNA regions; and functional constituents or domains thereof.

The DNA fragments according to the invention may moreover comprise regulatory sequences such as promoters, repressor or activator binding sites, repressor or activator genes, terminators; or structural genes. Likewise part of the invention are any combinations of these DNA fragments with one another or with other DNA fragments, for example combinations of promoters, repressor or activator binding sites and/or repressor or activator genes from an ansamycin gene cluster, in particular from the rifamycin gene cluster, with

foreign structural genes or combinations of structural genes from the ansamycin gene cluster, especially the rifamycin gene cluster, with foreign promoters; and combinations of structural genes with one another or with gene fragments which code for enzymatically active domains and are from various ansamycin biosynthesis systems. Foreign structural genes, and foreign gene fragments coding for enzymatically active domains, code, for example, for proteins involved in the biosynthesis of other ansamycins.

A preferred DNA fragment is one directly or indirectly involved in the gene cluster responsible for rifamycin synthesis.

The gene cluster or DNA region described above contains, for example, the genes which code for the individual enzymes involved in the biosynthesis of ansamycins and, in particular, of rifamycin, and the regulatory elements which control the expression of the biosynthesis genes. The size of such antibiotic gene clusters extends from about 10 kilobases (kb) up to over 100 kb. The gene clusters normally comprise specific regulatory genes and genes for resistance of the producer organism to its own antibiotic. Examples of what is meant by enzymes or enzymatically active domains involved in this biosynthesis are those necessary for synthesizing, starting from 3-amino-5-hydroxybenzoic acid, the ansamycins such as rifamycin, for example polyketide synthases, acyltransferases, dehydratases, ketoreductases, acyl carrier proteins or ketoacyl synthases.

Thus, the complete sequence of the gene cluster shown in SEQ ID NO 3, as well as DNA fragments which comprise sequence portions which code for a polyketide synthase or an enzymatically active domain thereof, are particularly preferred. Examples of such preferred DNA fragments are, for example, those which code for one or more of the proteins and polypeptides depicted in SEQ ID ID NOS 4, 5, 6, 7, 8 and 9, or functional derivatives thereof, also including partial sequences thereof which comprise, for example, 15 or more consecutive nucleotides. Other preferred embodiments relate to DNA regions of the gene cluster according to the invention or fragments thereof, like those present in the deposited clones pNE95, pRi44-2 and pNE112, or derived therefrom. Further preferred DNA fragments are those comprising sequence portions which display homologies with the sequences comprised by the clones pNE95, pRi44-2 and/or pNE112 or with SEQ ID ID NOS 1 and/or 3, and therefore can be used as hybridization probe within a genomic gene bank of an ansamycin-, in particular, rifamycin-producing organism for finding constituents

of the corresponding gene cluster. The DNA fragment may moreover, for example, comprise exclusively genomic DNA. A particularly preferred DNA fragment is one which comprises the nucleotide sequence depicted in SEQ ID NO 1 or 3, or partial sequences thereof, which, by reason of homologies, can be regarded as structural or functional equivalent to said sequence or partial sequence therefrom, and which therefore are able to hybridize with this sequence.

The DNA fragments according to the invention comprise, for example, sequence portions which comprise homologies with the above-described enzymes, enzyme domains or fragments thereof.

The term homologies and structural and/or functional equivalents refers primarily to DNA and amino acid sequences with few or minimal differences between the relevant sequences. These differences may have very diverse causes. Thus, for example, this may entail mutations or strain-specific differences which occur naturally or are artificially induced. Or the differences observed from the initial sequence are derived from a targeted modification, which can be introduced, for example, during a chemical synthesis.

Functional differences can be regarded as minimal if, for example, the nucleotide sequence coding for a polypeptide, or a protein sequence has essentially the same characteristic properties as the initial sequence, whether in respect of enzymatic activity, immunological reactivity or, in the case of a nucleotide sequence, gene regulation.

Structural differences can be regarded as minimal as long as there is a significant overlap or similarity between the various sequences, or they have at least similar physical properties. The latter include, for example, the electrophoretic mobility, chromatographic similarities, sedimentation coefficients, spectrophotometric properties etc.

In the case of nucleotide sequences, the agreement should be at least 70%, but preferably 80% and very particularly preferably 90% or more. In the case of the amino acid sequence, the corresponding figures are at least 50%, but preferably 60% and particularly preferably 70%. 90% agreement is very particularly preferred.

The invention furthermore relates to a method for identifying, isolating and cloning one of the DNA fragments described above. A preferred method comprises, for example, the following steps:

a) setting up of a genomic gene bank,

b) screening of this gene bank with the assistance of the DNA sequences according to the invention, and

c) isolation of the clones identified as positive.

A general method for identifying DNA fragments involved in the biosynthesis of ansamycins comprises, for example, the following steps

1) Cloning of a DNA fragment which shows homology with known polyketide synthase genes.

a) The presence of DNA fragments having homology with the polyketide synthase genes according to the invention is detected in the strains of the microorganism to be investigated by a Southern experiment with chromosomal DNA of this strain. The size of such homologous DNA fragments can be determined by digesting the DNA with a suitable restriction enzyme.

b) Production of a plasmid gene bank comprising the above digested chromosomal fragments. Normally, individual clones of this gene bank are tested once again for homology with the polyketide synthase genes according to the invention. Clones with recombinant plasmids comprising fragments having homology with the polyketide probe are then normally isolated on the basis of this homology.

2) Analysis of the cloned region

a) Restriction analysis of the isolated recombinant plasmids and checking of the identity of these cloned fragments with one another.

b) By a chromosomal Southern with DNA of the original microorganism and the isolated DNA fragment as probe it can be demonstrated that the cloned fragment is an original chromosomal DNA fragment from the original microorganism.

c) It is possible as an option to demonstrate a significant homology of the cloned DNA fragment with chromosomal DNA from other ansamycin producers (streptovaricin, tolypomycin, geldanamycin, ansamitocin). This would confirm that the cloned DNA is typical of gene clusters of ansamycin biosynthesis and thus also of rifamycin biosynthesis.

d) DNA sequencing of an internal restriction fragment and demonstration by comparative sequence analysis that the cloned region is a typical DNA sequence of polyketide synthases, coding for the biosynthesis of polyketide antibiotics from actinomycetes.

3) Isolation and characterization of adjacent DNA regions

a) Construction of a cosmid gene bank from the original microorganism and analysis thereof for homology with the isolated fragments. Isolation of cosmids having homology with this fragment.

b) Demonstration by restriction analysis that the isolated cosmid clones comprise a DNA region of the original microorganism which overlaps with the original fragment.

As described above, the first step in the isolation of the DNA fragments according to the invention is normally the setting up of genomic gene banks from the organism of interest, which synthesize the required ansamycin, especially rifamycin.

Genomic DNA can be obtained from a host organism in various ways, for example by extraction from the nuclear fraction and purification of the extracted DNA by known methods.

The fragmentation, which is necessary for setting up a representative gene bank, of the genomic DNA to be cloned to a size which is suitable for insertion into a cloning vector can take place either by mechanical shearing or else, preferably, by cutting with suitable restriction enzymes.

Suitable cloning vectors, which are already in routine use for producing genomic gene libraries, comprise, for example, cosmid vectors, plasmid vectors or phage vectors.

It is then possible in a screening program to obtain suitable clones which comprise the required gene(s) or gene fragment(s) from the gene libraries produced in this way.

One possibility for identifying the required DNA region consists in, for example, using the gene bank described above to transform strains which, because of a blocked synthetic pathway, are unable to produce ansamycins, and identifying those clones which are again able after the transformation to produce ansamycin (revertants). The vectors which lead to revertants comprise a DNA fragment which is required in ansamycin synthesis.

Another possibility for identifying the required DNA region is based, for example, on using suitable probe molecules (DNA probe) which are obtained for example as described above. Various standard methods are available for identifying suitable clones, such as differential colony hybridization or plaque hybridization.

It is possible to use as probe molecule a previously isolated DNA fragment from the same or a structurally related gene or gene cluster which, because of the homologies present, is able to hybridize with the corresponding sequence section within the required gene or gene cluster to be identified. Preferably used as probe molecule for the purpose of the present invention is a DNA fragment obtainable from a gene or a DNA sequence involved in the synthesis of polyketides such as ansamycins or soraphens.

If the nucleotide sequence of the gene to be isolated, or at least parts of this sequence, are known, it is possible in an alternative embodiment to use, based on this sequence information, a corresponding synthesized DNA sequence for the hybridizations or PCR amplifications.

In order to facilitate detectability of the required gene or else parts of a required gene, one of the DNA probe molecules described above can be labelled with a suitable, easily detectable group. A detectable group for the purpose of this invention means any material which has a particular, easily identifiable, physical or chemical property.

Particular mention may be made at this point of enzymatically active groups such as enzymes, enzyme substrates, coenzymes and enzyme inhibitors, furthermore fluorescent and luminescent agents, chromophores and radioisotopes such as ^{3}H, ^{35}S, ^{32}P, ^{125}I and ^{14}C. Easy detectability of these markers is based, on the one hand, on their intrinsic physical properties (for example fluorescent markers, chromophores, radioisotopes) or, on the other hand, on their reaction and binding properties (for example enzymes, substrates, coenzymes, inhibitors). Materials of these types are already widely used in particular in immunoassays and, in most cases, can also be used in the present application.

General methods relating to DNA hybridization are described, for example, by Maniatis T. *et al.*, Molecular Cloning, Cold Spring Harbor Laboratory Press (1982).

Those clones within the previously described gene libraries which are able to hybridize with a probe molecule and which can be identified by one of the abovementioned detection methods can then be further analysed in order to determine the extent and nature of the coding sequence in detail.

An alternative method for identifying cloned genes is based on constructing a gene library consisting of plasmid or expression vectors. This entails, in analogy to the methods described previously, the genomic DNA comprising the required gene being initially isolated and then cloned into a suitable plasmid or expression vector. The gene libraries produced in this way can then be screened by suitable procedures, for example by use of complementation studies, and those clones which comprise the required gene or else at least a part of this gene as insert can be selected.

It is thus possible with the aid of the methods described above to isolate a gene, several genes or a gene cluster which code for one or more particular gene products.

For further characterization, the DNA sequences purified and isolated in the manner described above are subjected to restriction analysis and sequence analysis.

For sequence analysis, the previously isolated DNA fragments are first fragmented using suitable restriction enzymes, and then cloned into suitable cloning vectors. In order to avoid mistakes in the sequencing, it is advantageous to sequence both DNA strands completely.

Various alternatives are available for analysing the cloned DNA fragment in respect of its function within ansamycin biosynthesis.

Thus, for example, it is possible in complementation experiments with defective mutants not only to establish involvement in principle of a gene or gene fragment in secondary metabolite biosynthesis, but also to verify specifically the synthetic step in which said DNA fragment is involved.

In an alternative type of analysis, evidence is obtained in exactly the opposite way. Transfer of plasmids which comprise DNA sections which have homologies with appropriate sections

on the genome results in integration of said homologous DNA sections via homologous recombination. If, as in the present case, the homologous DNA section is a region within an open reading frame of the gene cluster, plasmid integration results in inactivation of this gene by so-called gene disruption and, consequently, in an interruption in secondary metabolite production. It is assumed according to current knowledge that a homologous region which comprises at least 100 bp, but preferably more than 1000 bp, is sufficient to bring about the required recombination event.

However, a homologous region which extends over a range of from 0.3 to 4 kb, but in particular over a range of from 1 to 3 kb, is preferred.

To prepare suitable plasmids which have sufficient homology for integration via homologous recombination there is preferably provision of a subcloning step in which the previously isolated DNA is digested, and fragments of suitable size are isolated and subsequently cloned into a suitable plasmid. Examples of suitable plasmids are the plasmids generally used for genetic manipulations in streptomycetes or *E. coli*.

It is possible in principle to use for the preparation and multiplication of the previously described constructs all conventional cloning vectors such as plasmid or bacteriophage vectors as long as they have replication and control sequences derived from species compatible with the host cell.

The cloning vector usually has an origin of replication plus specific genes which result in phenotypical selection features in the transformed host cell, in particular resistances to antibiotics. The transformed vectors can be selected on the basis of these phenotypical markers after transformation in a host cell.

Selectable phenotypical markers which can be used for the purpose of this invention comprise, for example, without this representing a limitation of the subject-matter of the invention, resistances to thiostrepton, ampicillin, tetracycline, chloramphenicol, hygromycin, G418, kanamycin, neomycin and bleomycin. Another selectable marker can be, for example, prototrophy for particular amino acids.

Mainly preferred for the purpose of the present invention are streptomycetes and *E. coli* plasmids, for example the plasmids used for the purpose of the present invention.

Host cells primarily suitable for the previously described cloning for the purpose of this invention are prokaryotes, including bacterial hosts such as streptomycetes, actinomycetes, *E. coli* or pseudomonads.

E. coli hosts are particularly preferred, for example the *E. coli* strain HB101 or X-1 blue MR®(Stratagene) or streptomyces such as the plasmid-free strains of *Streptomyces lividans* TK23 and TK24.

Competent cells of the *E. coli* strain HB101 are produced by the methods normally used for transforming *E. coli*. The transformation method of Hopwood *et al.* (Genetic manipulation of streptomyces a laboratory manual. The John Innes Foundation, Norwich (1985)) is normally used for streptomyces.

After transformation and subsequent incubation on a suitable medium, the resulting colonies are subjected to a differential screening by plating out on selective media. It is then possible to isolate the appropriate plasmid DNA from those colonies which comprise plasmids with DNA fragments cloned in.

The DNA fragment according to the invention, which comprises a DNA region which is involved directly or indirectly in the biosynthesis of ansamycin and can be obtained in the previously described manner from the ansamycin biosynthesis gene cluster, can also be used as starter clone for identifying and isolating other adjacent DNA regions overlapping therewith from said gene cluster.

This can be achieved, for example, by carrying out a so-called chromosome walking within a gene library consisting of DNA fragments with mutually overlapping DNA regions, using the previously isolated DNA fragment or else, in particular, the sequences located at its 5' and 3' margins. The procedures for chromosome walking are known to the person skilled in this art. Details can be found, for example, in the publications by Smith *et al.* (Methods

Enzymol (1987), **151**, 461-489) and Wahl *et al.* (Proc Natl. Acad. Sci, USA (1987), **84**, 2160-2164).

The prerequisite for chromosome walking is the presence of clones having coherent DNA fragments which are as long as possible and mutually overlap within a gene library, and a suitable starter clone which comprises a fragment which is located in the vicinity or else, preferably, within the region to be analysed. If the exact location of the starter clone is unknown, the walking is preferably carried out in both directions.

The actual walking step starts by using the identified and isolated starter clone as probe in one of the previously described hybridization reactions in order to detect adjacent clones which have regions overlapping with the starter clone. It is possible by hybridization analysis to establish which fragment projects furthest over the overlapping region. This is then used as starting clone for the 2nd walking step, in which case there is establishment of the fragment which overlaps with said 2nd clone in the same direction. Continuous progression in this manner on the chromosome results in a collection of overlapping DNA clones which cover a large DNA region. These can then, where appropriate after one or more subcloning steps, be ligated together by known methods to give a fragment which comprises parts or else, preferably all of the constituents essential for ansamycin biosynthesis.

The hybridization reaction to establish clones with overlapping marginal regions preferably makes use not of the very large and unwieldy complete fragment but, in its place, a partial fragment from the left or right marginal region, which can be obtained by a subcloning step. Because of the smaller size of said partial fragment, the hybridization reaction results in fewer positive hybridization signals, so that the analytical effort is distinctly less than on use of the complete fragment. It is furthermore advisable to characterize the partial fragment in detail in order to preclude its comprising larger amounts of repetitive sequences, which may be distributed over the entire genome and thus would greatly impede a targeted sequence of walking steps.

Since the gene cluster responsible for ansamycin biosynthesis covers a relatively large region of the genome, it may also be advantageous to carry out a so-called large-step walking or cosmid walking. It is possible in these cases, by using cosmid vectors which

permit the cloning of very large DNA fragments, to cover a very large DNA region, which may comprise up to 42 kb, in a single walking step.

In one possible embodiment of the present invention, for example, to construct a cosmid gene bank from streptomycetes or actinomycetes, complete DNA is isolated with the size of the DNA fragments being of the order of about 100 kb, and is subsequently partially digested with suitable restriction endonucleases.

The digested DNA is then extracted in a conventional way in order to remove endonuclease which is still present, and is precipitated and finally concentrated. The resulting fragment concentrate is then fractionated, for example by density gradient centrifugation, in accordance with the size of the individual fragments. After the fractions obtainable in this way have been dialysed they can be analysed on an agarose gel. The fractions which contain fragments of suitable size are pooled and concentrated for further processing. Fragments to be regarded as particularly suitable for the purpose of this invention have a size of the order of 30 kb to 42 kb, but preferably of 35 kb to 40 kb.

In parallel with the fragmentation described above, or later, for example a suitable cosmid vector pWE15® (Stratagene) is completely digested with a suitable restriction enzyme, for example BamHI, for the subsequent ligase reaction.

Ligation of the cosmid DNA to the streptomyces or actinomycetes fragments which have been fractionated according to their size can be carried out using a T4 DNA ligase. The ligation mixture obtainable in this way is, after a sufficient incubation time, packaged into λ phages by generally known methods.

The resulting phage particles are then used to infect a suitable host strain. A recA⁻ *E. coli* strain is preferred, such as *E. coli* HB101 or X-1 Blue® (Stratagene). Selection of transfected clones and isolation of the plasmid DNA can be carried out by generally known methods.

The screening of the gene bank for DNA fragments which are involved in ansamycin biosynthesis is carried out, for example, using a specific hybridization probe which is assumed (for example on the basis of DNA sequence or DNA homology or

complementation tests or gene disruption or the function thereof in other organisms) to comprise DNA regions from the 'ansamycin gene cluster'.

A plasmid which comprises an additional fragment of the required size or has been identified on the basis of hybridizations can then be isolated from the gel in the previously described manner. The identity of this additional fragment with the required fragment of the previously selected cosmid can then be confirmed by Southern transfer and hybridization.

Function analysis of the DNA fragments isolated in this way can be carried out in a gene disruption experiment as described above.

Another possible use of the DNA fragments according to the invention is to modify or inactivate enzymes or domains involved in ansamycin and, in particular, rifamycin biosynthesis, or to synthesize oligonucleotides which are then in turn used for finding homologous sequences in PCR amplification.

Besides the DNA fragments according to the invention as such, also claimed are their use firstly for producing rifamycin, rifamycin analogues or precursors thereof, and for the biosynthetic production of novel ansamycins or of precursors thereof. Included in this connection are those molecules in which the aliphatic bridge is connected only at one end to the aromatic nucleus.

The DNA fragments according to the invention permit, for example, by combination with DNA fragments from other biosynthetic pathways or by inactivation or modification thereof, the biosynthesis of novel hybrid compounds, in particular of novel ansamycins or rifamycin analogues. The steps necessary for this are generally known and are described, for example, in Hopwood, Current Opinion in Biotechnol. (1993), 4, 531-537.

The invention furthermore relates to the use of the DNA fragments according to the invention for carrying out the novel technology of combinatorial biosynthesis for the biosynthetic production of libraries of polyketide synthases based on the rifamycin and ansamycin biosynthesis genes. If, for example, several sets of modifications are produced, it is possible in this way to produce, by means of biosyntheses, a library of polyketides, for example ansamycins or rifamycin analogues, which then needs to be tested only for the

activity of the compounds produced in this way. The steps necessary for this are generally known and are described, for example, in Tsoi and Khosla, Chemistry & Biology (1995), 2, 355-362 and WO-9508548.

Besides the DNA fragment as such, also claimed is its use for the genetic construction of mutated actinomycetes strains from which the natural rifamycin or ansamycin biosynthesis gene cluster in the chromosome has been partly or completely deleted, and which can thus be used for expressing genetically modified ansamycin or rifamycin biosynthesis gene clusters.

The invention furthermore relates to a hybrid vector which comprises at least one DNA fragment according to the invention, for example a promoter, a repressor or activator binding site, a repressor or activator gene, a structural gene, a terminator or a functional part thereof. The hybrid vector comprises, for example, an expression cassette which comprises a DNA fragment according to the invention which is able to express one or more proteins involved in ansamycin biosynthesis and, in particular in rifamycin biosynthesis, or a functional fragment thereof. The invention likewise relates to a host organism which comprises the hybrid vector described above.

Suitable vectors representing the starting point of the hybrid vectors according to the invention, and suitable host organisms such as bacteria or yeast cells are generally known.

The host organism can be transformed by generally customary methods such as by means of protoplasts, Ca^{2+}, Cs^{+}, polyethylene glycol, electroporation, viruses, lipid vesicles or a particle gun. The DNA fragments according to the invention may then be present both as extrachromosomal constituents in the host organism and integrated via suitable sequence sections into the chromosome of the host organism.

The invention likewise relates to polyketide synthases which comprise the DNA fragments according to the invention, in particular those from *Amycolatopsis mediterranei* which are involved directly or indirectly in rifamycin synthesis, and functional constituents thereof, for example enzymatically active domains.

The invention furthermore relates to a hybridization probe comprising a DNA fragment according to the invention, and to the use thereof, in particular for identifying DNA fragments involved in the biosynthesis of ansamycins.

In order to obtain unambiguous signals in the hybridization, DNA bound to the filter (for example made of nylon or nitrocellulose) is normally washed at 55-65°C in 0.2 × SSC (1 × SSC = 0.15 M sodium chloride, 15 mM sodium citrate).

Examples

General

General molecular genetic techniques such as DNA isolation and purification, restriction digestion of DNA, agarose gel electrophoresis of DNA, ligation of restriction fragments, cultivation and transformation of *E. coli*, plasmid isolation from *E. coli*, are carried out as described in Maniatis et al., Molecular Cloning: A laboratory manual, 1st Edit. Cold Spring Harbor Laboratory Press, Cold Spring Harbor NY (1982).

Culture conditions and molecular genetic techniques with *A. mediterranei* and other *actinomycetes* are as described by Hopwood *et al.* (Genetic manipulation of streptomyces a laboratory manual, The John Innes Foundation, Norwich, 1985). All liquid cultures of *A. mediterranei* and other *actinomycetes* are carried out in Erlenmeyer flasks at 28°C on a shaker at 250 rpm.

Nutrient media used:

LB Maniatis et al., Molecular Cloning: A laboratory manual, 1st Edit. Cold Spring Harbor Laboratory Press, Cold Spring Harbor NY (1982)

NL148 Schupp + Divers FEMS Microbiology Lett. **36**, 159-162 (1986) (NL148 = NL148G without glycine)

R2YE Hopwood *et al.* (Genetic manipulation of streptomyces a laboratory manual, The John Innes Foundation, Norwich, 1985)

TB : 12 g/l Bacto tryptone
24 g/l Bacto yeast extract
4 ml/l glycerol

Example 1: Detection of chromosomal DNA fragments from *A. mediterranei* having homology with polyketide synthase genes of other bacteria

To obtain genomic DNA from *A. mediterranei*, cells of the strain *A. mediterranei* wt3136 (= LBGA 3136, ETH collection of strains) are cultivated in NL148 medium for 48 hours. 1 ml of this culture is then transferred into 50 ml of NL148 medium (+ 2.5 g/l glycine) in a 200 ml Erlenmeyer flask, and the culture is incubated for 48 h. The cells are removed from the medium by centrifugation at 3000 g for 10 min. and are resuspended in 5 ml of SET (75 mM NaCl, 25 mM EDTA, 20 mM Tris, pH 7.5). High molecular weight DNA is extracted by the method of Pospiech and Neumann (Trends in Genetics (1995), 11, 217-218).

In order to detect, by a Southern blot, individual fragments from the isolated *A. mediterranei* DNA which have homology with polyketide synthase genes, a radioactive DNA probe is prepared from a known polyketide synthase gene cluster. To do this, the PvuI fragment 3.8 kb in size is isolated from the recombinant plasmid p98/1 (Schupp et al. J. of Bacteriol. (1995), 177, 3673-3679), which comprises a DNA region, about 32 kb in size, from the polyketide synthase for the antibiotic soraphen A. About 0.5 µg of the isolated 3.8 kb PvuI DNA fragment is radiolabelled with ^{32}P-d-CTP by the nick translation system from Gibco/BRL (Basle) in accordance with the manufacturer's instructions.

For the Southern blot, about 2 µg of the genomic DNA isolated above from *A. mediterranei* are completely digested with the restriction enzyme BglII (Böhringer, Mannheim), and the resulting fragments are fractionated on a 0.8% agarose gel. A Southern blot with this agarose gel and the DNA probe isolated above (3.8 kb PvuI fragment) detects a DNA BglII-cut fragment which is about 13 kb in size from the genomic DNA of *A. mediterranei*, and which has homology with the DNA probe used. It can be concluded on the basis of this homology that the detected DNA fragment from *A. mediterranei* is a genetic region which codes for a polyketide synthase and thus is involved in the synthesis of a polyketide antibiotic.

Example 2: Production of a specific recombinant plasmid collection comprising BglII-digested chromosomal fragments from *A. mediterranei* 12-16 kb in size

The *E. coli* positive selection vector pIJ4642 (derivative of pIJ666, Kieser & Melton, Gene (1988), 65, 83-91) developed at the John Innes Centre (Norwich, UK) is used to produce the plasmid gene bank. This plasmid is first cut with BamHI, and the two resulting fragments are fractionated on an agarose gel. The smaller of the two fragments is the filler fragment of the vector and the larger is the vector portion which, on self-ligation after deletion of the filler fragment, forms, owing to the flanking fd termination sequences, a perfect palindrome, which means that the plasmid cannot be obtained as such in *E. coli*. This vector portion 3.8 kb in size is isolated from the agarose gel by electroelution as described on page 164-165 of Maniatis et al., Molecular Cloning: A laboratory manual, 1st Edit. Cold Spring Harbor Laboratory Press, Cold Spring Harbor NY (1982).

To prepare the BglII-cut DNA fragments from *A. mediterranei*, the high molecular weight genomic DNA prepared in Example 1 is used. About 10 µg of this DNA are completely digested with the restriction enzyme BglII and subsequently fractionated on a 0.8% agarose gel. DNA fragments with a size of about 12 - 16 kb are cut out of the gel and detached from the gel block by electroelution (see above). About 1 µg of the BglII fragments isolated in this way is ligated to about 0.1 µg of the BamHI portion, isolated above, of the vector pIJ4642. The ligation mixture obtained in this way is then transformed into the *E. coli* strain HB101 (Stratagene). About 150 transformed colonies are selected from the transformation mixture on LB agar with 30 µg per ml chloramphenicol. These colonies contain recombinant plasmids with BglII-cut genomic DNA fragments from *A. mediterranei* in the size range 12 - 16 kb.

Example 3: Cloning and characterization of chromosomal *A. mediterranei* DNA fragments having homology with bacterial polyketide synthase genes

150 of the plasmid clones prepared in Example 2 are analysed by colony hybridization using a nitrocellulose filter (Schleicher & Schuell) as described on pages 318-319 of Maniatis et al., Molecular Cloning: A laboratory manual, 1st Edit. Cold Spring Harbor Laboratory Press, Cold Spring Harbor NY (1982). The DNA probe used is the 3.8 kb PvuI fragment, radiolabelled with ^{32}P-d-CTP and isolated in Example 1, of the plasmid p98/1. The plasmids are isolated from 5 plasmid clones which show a hybridization signal, and are characterized by two restriction digestions with the enzymes HindIII or KpnI. HindIII cuts

twice in the vector portion of the clones, 0.3 kb to the right and left of the BamHI cleavage site into which the *A. mediterranei* DNA has been integrated. KpnI does not cut in the pIJ 4642 vector portion. This restriction analysis shows that the investigated clones comprise both identical HindIII fragments of about 14 and 3.1 kb and identical KpnI fragments approximately 11.4 kb and 5.7 kb in size. This shows that these clones comprise the same genomic BglII fragment of *A. mediterranei*, and that the latter has a size of about 13 kb. It can additionally be concluded from this restriction analysis that this cloned BglII fragment has no internal HindIII cleavage site, but has 2 KpnI cleavage sites which afford an internal KpnI fragment 5.7 kb in size.

The plasmid DNA of the above 5 clones with identical restriction fragments is further characterized by a Southern blot. For this purpose, the plasmids are cut with HindIII and KpnI, and the DNA probe used is the ^{32}P-radiolabelled 3.8 kb PvuI fragment of the plasmid p98/1 used above. This experiment confirms that the 5 plasmids contain identical *A. mediterranei* DNA fragments and that these have significant homology with the DNA probe which is characteristic of bacterial polyketide synthase genes. In addition, the Southern blot shows that the internal KpnI fragment 5.7 kb in size likewise has significant homology with the DNA probe used. The plasmid called pRi7-3 is selected from the 5 plasmids for further processing.

To demonstrate that the cloned BglII fragment about 13 kb in size from *A. mediterranei* is an original chromosomal DNA fragment, another Southern blot is carried out. Chromosomal DNA from *A. mediterranei* which has been cut with BglII, KpnI or BamHI is employed in this blot. Two BamHI fragments which are about 1.8 and 1.9 kb in size and are present in the 5.7 kb KpnI fragment of pRi7-3 are used as radiolabelled DNA probe. This experiment confirms that the BglII DNA fragment about 13 kb in size cloned in the recombinant plasmid pRi7-3 is an authentic genomic DNA fragment from *A. mediterranei*. In addition, this experiment confirms that the cloned fragment comprises an internal KpnI fragment 5.7 kb in size and two BamHI fragments about 1.8 and 1.9 kb in size, and that these DNA fragments are likewise authentic genomic DNA fragments from *A. mediterranei*.

Example 4: Demonstration of a significant homology of the cloned genomic 13 kb BglII fragment from *A. mediterranei* with chromosomal DNA from other *actinomycetes* which produce ansamycins

Demonstration of a significant homology between the cloned chromosomal DNA region of *A. mediterranei* and chromosomal DNA from other ansamycin-producing actinomycetes takes place by a Southern blot experiment. The following ansamycin-producing strains are employed for this purpose (the ansamycins produced by the strains are in parentheses): *Streptomyces spectabilis* (streptovaricins), *Streptomyces tolypophorus* (tolypomycins), *Streptomyces hygroscopicus* (geldanamycins), *Nocardia species* ATCC31281 (ansamitocins). Genomic DNA from these strains is isolated as described for *A. mediterranei* in Example 1 and digested with the restriction enzyme KpnI, and the restriction fragments obtained in this way are fractionated on an agarose gel for the Southern blot. Two BamHI fragments about 1.8 and 1.9 kb in size from *A. mediterranei*, which are used in Example 3 and are isolated from the plasmid pRi7-3, are used as radioactive probe. This experiment shows that these ansamycin-producing strains have a significant DNA homology with the DNA probe used and thus with the cloned chromosomal region of *A. mediterranei*. It is to be observed in this connection that the homology in the case of producers of ansamycins with a naphthoquinoid ring system (streptovaricin, tolypomycin) is greater than in the case of those with a benzoquinoid ring system (geldanamycin, ansamitocin). This result suggests that the cloned chromosomal DNA region from *A. mediterranei* is typical of ansamycin biosynthesis gene clusters and, especially, of gene clusters for ansamycins with naphthoquinoid ring systems, corresponding to the ring system in rifamycins.

Example 5: DNA sequence determination of the KpnI fragment 5.7 kb in size located within the cloned 13 kb BglII fragment

For the sequencing, the 5.7 kb KpnI fragment is isolated from the plasmid pRi7-3 (DSM 11114) (Maniatis et. al. 1992) and subcloned into the KpnI cleavage site of the vector pBRKanf4, which is suitable for the DNA sequencing, affording the plasmids pTS004 and pTS005. The vector pBRKanf4 (derived from pBRKanf1; Bhat, Gene (1993) **134**, 83-87) is suitable for introducing sequential deletions of Sau3A fragments in the cloned insert fragment, because this vector does not itself have a GATC nucleotide sequence. In addition, the BamHI fragments 1.9 and 1.8 kb in size present in the 5.7 kb KpnI fragment are subcloned into the BamHI cleavage site of pBRKanf4, resulting the plasmids pTS006 and pTS007, and pTS008 and pTS009, respectively.

To prepare subclones sequentially truncated by Sau3A fragments for the DNA sequencing, the plasmids pTS004 to pTS009 are partially digested with Sau3A and completely digested with XbaI or HindIII (a cleavage site in the multiple cloning region of the vector). The DNA obtained in this way (consisting of the linearized vector with inserted DNA fragments truncated by Sau3A fragments) is filled in at the ends using Klenow polymerase (fragment of polymerase I, see Maniatis *et al.* pages 113-114), self-ligated with T4 DNA ligase and transformed into *E. coli* DH5α. The plasmid DNA which corresponds to the pTS004 to pTS009 plasmids, but has DNA regions, which are truncated from one side by Sau3A fragments, from the original integrated fragments of *A. mediterranei*, is isolated from individual transformed clones obtained in this way.

The DNA sequencing is carried out with the plasmids obtained in this way and with pTS004 to pTS009 using the reaction kit from Perkin-Elmer/Applied Biosystems with dye-labelled terminator reagents (Kit N° 402122) and a universal primer or a T7 primer. A standard cycle sequencing protocol with a thermocycler (MJ Research DNA Engine Thermocycler, Model 225) is used, and the sequencing reactions are analysed by the Applied Biosystems automatic DNA sequencer (Modell 373 or 377) in accordance with the manufacturer's instructions. To analyse the results, the following computer programs (software) are employed: Applied Biosystems DNA analysis software, Unix Solaris CDE software, DNA assembly and analysis package GAP licensed from R. Staden (Nucleic Acid Research (1995)23, 1406-1410) and Blast (NCBI).

The methods described above can be used to sequence completely both DNA strands of the 5.7 kb KpnI fragment from *A. mediterranei* strain wt3136. The DNA sequence of the 5.7 kb fragment with a length of 5676 base pairs is depicted in SEQ ID NO 1.

Example 6: Analysis of the protein-encoding region (genes) on the 5.7 kb KpnI fragment from *A. mediterranei*

The nucleotide sequence of the 5.7 kb KpnI fragment is analysed using the Codonpreference computer program (Genetics Computer Group, University of Wisconsin, 1994). This analysis shows that this fragment is over its whole length a protein-encoding region and thus forms part of a larger open reading frame (ORF). The codons used in this ORF are typical of

streptomycetes and actinomycetes genes. The amino acid sequence derived from the DNA sequence from this ORF is depicted in SEQ ID NO 2.

Polyketide synthases for macrolide antibiotics (such as erythromycin, rapamycin) are very large multifunctional proteins which comprise several enzymatically active domains which are now well characterized (Hopwood und Khosla, Ciba Foundation Symposium (1992), 171, 88-112; Donadio and Katz, Gene (1992), 111, 51-60; Schwecke et al., Proc. Natl. Acad. Sci. U.S.A. (1995) 92 (17), 7839-7843). Comparison of the amino acid sequence depicted in SEQ ID NO 2 with that of the very well-characterized erythromycin polyketide synthase, eryA ORF1 (Donadio, Science, (1991) 252, 675-679, DNA sequence gene/EMBL accession NO M63676) gives the following results:

Region from SEQ ID NO 2: amino acids 2 - 325: is 40% identical to the acyltransferase domain of module 2 of the *eryA* locus of *Saccharopolyspora erythraea*.

Region from SEQ ID NO 2: amino acids 325 - 470: is 43% identical to the dehydratase domain of module 4 of the *eryA* locus of *Saccharopolyspora erythraea*.

Region from SEQ ID NO 2: amino acids 762 - 940: is 46% identical to the ketoreductase domain of module 2 of the *eryA* locus of *Saccharopolyspora erythraea*.

Region from SEQ ID NO 2: amino acids 1024- 1109: is 57% identical to the acyl carrier protein domain of module 2 of the *eryA* locus of *Saccharopolyspora erythraea*.

Region from SEQ ID NO 2: amino acids 1126 - 1584: is 59% identical to the ketoacyl synthase domain of module 1 of the *eryA* locus of *Saccharopolyspora erythraea*.

The very large similarities found in the amino acid sequence and in the size and arrangement of the enzymatic domains suggest that the cloned KpnI region 8.7 kb in size from *A. mediterranei* codes for part of a polyketide synthase which is typical of polyketides of the macrolide type.

Example 7: Construction of a cosmid gene bank from *A. mediterranei*

The cosmid vector employed is the plasmid pWE15 which can be purchased (Stratagene, La Jolla, CA, USA). pWE15 is completely cut with the enzyme BamHI (Maniatis *et al.* 1989) and precipitated with ethanol. For ligation to the cosmid DNA, chromosomal DNA from *A. mediterranei* is isolated as described in Example 1 and partially digested with the restriction enzyme Sau3A (Böhringer, Mannheim) to form DNA fragments most of which have a size of 20 - 40 kb. The DNA pretreated in this way is fractionated by fragment size by centrifugation (83,000 g, 20°C) on a 10% to 40% sucrose density gradient for 18 h. The gradient is fractionated in 0.5 ml aliquots and dialysed, and samples of 10 µl are analysed on a 0.3% agarose gel with DNA size standard. Fractions with chromosomal DNA 25 - 40 kb in size are combined, precipitated with ethanol and resuspended in a small volume of water.

Ligation of the cosmid DNA to the *A. mediterranei* Sau3A fragments isolated according to their size (see above) takes place with the aid of a T4-DNA ligase. About 3 µg of each of the two DNA starting materials are employed in a reaction volume of 20 µl, and the ligation is carried out at 12°C for 15 h. 4 ml of this ligation mixture are packaged into lambda phages using the *in vitro* packaging kit which can be purchased from Stratagene (La Jolla, CA, USA) (in accordance with the manufacturer's instructions). The resulting phages are introduced by infection into the *E. coli* strain X-1BlueMR® (Stratagene). Titration of the phage material reveals about 20,000 phage particles per ml, analysis of 12 cosmid clones shows that all the clones contain plasmid DNA inserts 25 - 40 kb in size.

Example 8: Identification, cloning and characterization of the chromosomal *A. mediterranei* DNA region which is adjacent to the cloned 5.7 kb KpnI fragment

To identify and clone the chromosomal *A. mediterranei* DNA region which is adjacent to the 5.7 kb KpnI fragment described above in Examples 3 and 5, firstly a radioactive DNA probe is prepared from this 5.7 kb KpnI fragment. This is done by radiolabelling approximately 0.5 µg of the isolated DNA fragment with ^{32}P-d-CTP by the nick translation system of Gibco/BRL (Basle) in accordance with the manufacturer's instructions.

Infection of *E. coli* X-1 Blue MR (Stratagene) with an aliquot of the lambda phages packaged *in vitro* (see Example 7) results in more than 2000 clones on several LB + ampicillin (50 µg/ml) plates. These clones are tested by colony hybridization on nitrocellulose filters (see Example 3 for method). The DNA probe used is the 5.7 kb KpnI DNA fragment from *A. mediterranei* which is radiolabelled with ^{32}P-d-CTP and was prepared above.

5 cosmid clones showing a significant signal with the DNA probe are found. The plasmid DNA of these cosmids is isolated (Sambrook et al., Molecular Cloning: A Laboratory Manual, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, 1989), digested with KpnI and analysed in an agarose gel. Analysis reveals that all 5 plasmids have integrated chromosomal *A. mediterranei* DNA with a size of the order of about 25-35 kb, and all contain the 5.7 kb KpnI fragment.

To characterize the chromosomal *A. mediterranei* DNA region which is adjacent to the cloned KpnI fragment, the plasmid DNA of one of the 5 cosmid clones is subjected to restriction analysis. The selected plasmid of the cosmid clone has the number pNE112 and likewise comprises the 13 kb BglII fragment described in Example 3.

Digestion of the plasmid pNE112 with the restriction enzymes BamHI, BglII, HindIII (singularly and in combination) allows a restriction map of the cloned region of *A. mediterranei* to be prepared, and this permits this region about 26 kb in size in the chromosome of *A. mediterranei* to be characterized. This region is characterized by the following restriction cleavage sites with the stated distance in kb from one end: BamHI in position 3.2 kb, HindIII in position 6.6 kb, BglII in position 11.5 kb, BamHI in position 16.6 kb, BamHI in position 17.3 kb, BamHI in position 21 kb and BglII in position 24 kb.

Example 9: Determination of the sequence of the chromosomal *A. mediterranei* DNA region present in the plasmid pNE112 and overlapping with the cloned 5.7 kb KpnI fragment

The plasmid pNE112 DNA is split up into fragments directly using an Aero-Mist nebulizer (CIS-US Inc., Bedford, MA, USA) under a nitrogen pressure of 8-12 pounds per square inch. These random DNA fragments are treated with T4 DNA polymerase, T4 DNA kinase and E. coli DNA polymerase in the presence of the 4 dNTPs in order to generate blunt ends

on the double-stranded DNA fragments (Sambrook et al., Molecular Cloning: A Laboratory Manual, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, 1989). The fragments are then fractionated in 0.8% low melting agarose (FMC SeaPlaque Agarose, Catalogue N° 50113), and fragments 1.5-2 kb in size are extracted by hot phenol extraction (Sambrook et al., Molecular Cloning: A Laboratory Manual, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, 1989). The DNA fragments obtained in this way are then ligated with the aid of T4 DNA ligase to the plasmid vector pBRKanf4 (see Example 5) or pBlueScript KS+ (Stratagene, La Jolla, CA, USA), each of which is cut once with square ends by appropriate restriction digestion (SmaI for pBRKanf4 and EcoRV for pBlueScript KS+), and is dephosphorylated on the ends by a treatment with alkaline phosphatase (Böhringer, Mannheim). The ligation mixture is then transformed into E. coli DH5α, and the cells are incubated overnight on LB agar with the appropriate antibiotic (kanamycin 40 µg/ml for pBRKanf4, ampicillin 100 µg/ml for pBlueScript KS+). Grown colonies are transferred singly into 1.25 ml of liquid TB medium with antibiotic in 96-well plates with wells of a volume of 2 ml, and incubated at 37°C overnight. Template DNA for the sequencing is prepared directly from these cultures by alkaline lysis (Birnboim, Methods in Enzymology (1983) 100, 243-255). The DNA sequencing takes place using the Perkin Elmer/Applied Biosystems reaction kit with dye-labelled terminator reagents (Kit N° 402122) and universal M13 mp18/19 primers or T3, T7 primers, or with primers prepared by us which bind to internal sequences. A standard cycle sequencing protocol with 20 cycles is used with a thermocycler (MJ Research DNA Engine Thermocycler, Model 225). The sequencing reactions are precipitated with ethanol, resuspended in formamide loading buffer and fractionated and analysed by electrophoresis using the Applied Biosystems automatic DNA sequencer (Model 377) in accordance with the manufacturer's instructions. Sequence files are produced with the aid of the Applied Biosystems DNA Analysis Software computer program and transferred to a SUN UltraSpark computer for further analysis. The following computer programs (software) are employed for analysing the results: DNA assembly and analysis package GAP (Genetics Computer Group, University of Wisconsin, R. Staden, Cambridge University UK) and the four programs: Phred, Cross-match, Phrad and Consed (P. Green, University of Washington, B. Ewing and D. Gordon, Washington University in Saint Louis). After the original sequences have been connected together to give longer coherent sequences (contigs), missing DNA sections are specifically sequenced with the aid of new primers (binding to sequenced sections), or by longer sequencing or sequencing the other strand.

It is possible with the method described above to sequence the entire chromosomal DNA region 26 kb in size from *A. mediterranei* which is cloned in pNE112. The DNA sequence is depicted in SEQ ID NO 3 in the base pair 27801 - 53789 section. The DNA sequence of the 5.7 kb KpnI fragment described in Example 5 is present in pNE112, and is depicted in SEQ ID NO 3 in the base pair 43093 - 48768 region.

Example 10: Identification and characterization of cosmid clones with chromosomal DNA fragments from *A. mediterranei* which overlap with one end of the 26 kb *A. mediterranei* region of pNE112

To identify cosmid clones which comprise chromosomal DNA fragments from *A. mediterranei* located directly in front of the 26 kb region of pNE112, the plasmid pNE112 is cut with the restriction enzyme BamHI, and the resulting BamHI fragment 3.2 kb in size is separated from the other BamHI fragments in an agarose gel and isolated from the gel. This BamHI fragment is located at one end of the incorporated *A. mediterranei* DNA in pNE112 (see Example 8) and can thus be used as DNA probe for finding the required cosmid clones. Approximately 0.5 µg of the isolated 3.2 kb BamHI DNA fragment is radiolabelled with ^{32}P-dCTP by the nick translation system from Gibco/BRL (Basel) in accordance with the manufacturer's instructions.

The cosmid gene bank from *A. mediterranei* described in Example 7 is then analysed by colony hybridization (Method of Example 3) using this 3.2 kb DNA probe for clones with overlaps. Two cosmid clones with a strong hybridization signal can be identified in this way and are given the numbers pNE95 and pRi44-2. It is possible by restriction analysis and Southern blot to confirm that the plasmids pNE95 and pRi44-2 comprise chromosomal DNA fragments from *A. mediterranei* which overlap with the 3.2 kb BamHI fragment from pNE112 and together cover a 35 kb chromosomal region of *A. mediterranei* which is directly adjacent to the 26 kb *A. mediterranei* fragment of pNE112 cloned in pNE112.

Example 11: Restriction analysis of the chromosomal *A. mediterranei* DNA region cloned with the cosmid clones pNE112, pNE95 and pRi44-2

The chromosomal *A. mediterranei* DNA region cloned with the cosmid clones pNE112, pNE95 and pRi44-2 is characterized by carrying out a restriction analysis. Digestion of the plasmid DNA of the three cosmids with the restriction enzymes EcoRI, BglII and HindIII (singly and in combination) produces a rough restriction map of the cloned region of *A. mediterranei*. Overlapping fragments of the three plasmids are in this case established and confirmed by Southern blot. This chromosomal region of *A. mediterranei* has a size of about 61 kb and is characterized by the following restriction cleavage sites with the stated distance in kb from one end: EcoRI in position 7.2 kb, HindIII in position 21 kb, BglII in position 31 kb, HindIII in position 42 kb, BglII in position 47 kb and BglII in position 59 kb. In this region in the *A. mediterranei* chromosome, the plasmid pRi 44-2 covers a region from position 1 to approximately 37 kb, plasmid pNE95 covers a region of approximate position 9 kb - 51 kb and plasmid pNE 112 covers a region of approximate position 35 kb - 61 kb.

Example 12: Determination of the sequence of the chromosomal *A. mediterranei* DNA region described in Example 11 from the EcoRI cleavage site in the 7.2 kb position up to the 61 kb end

Determination of the DNA sequence of the chromosomal region described in Example 11 from *A. mediterranei* (EcoRI cleavage site in the 7.2 kb position to 51 kb) is carried out with the plasmids pRi 44-2 and pNE95, using exactly the same method as described in Example 9. Analysis of the DNA sequence obtained in this way confirms the rough restriction map described in Example 11 and the overlaps of the cloned *A. mediterranei* fragments in the plasmids pNE112, pNE95 and pRi44-2.

The DNA sequence of the chromosomal *A. mediterranei* DNA region described in Example 11 from the EcoRI cleavage site in the 7.2 kb position up to the end at 61 kb is depicted in SEQ ID NO 3 (length 53789 base pairs).

Example 13: Analysis of a first protein-encoding region (ORF A) of the cloned *A. mediterranei* chromosomal region depicted in SEQ ID NO 3

The nucleotide sequence shown in SEQ ID NO 3 is analysed with the Codonpreference computer program (Genetics Computer Group, University of Wisconsin, 1994). This analysis shows that a very large open reading frame (ORF A) which codes for a protein is present in

the first third of the sequence (position 1825 - 15543 including stop codon in SEQ ID NO 3). The codons used in ORF A are typical of actinomycetes genes with a high G+C content.

Comparison of the amino acid sequence of ORF A (SEQ ID NO 4, size 4572 amino acids) with other polyketide synthases and specifically with the very well characterized polyketide synthase of *Saccharopolyspora erythraea* (Donadio, Science, (1991) 252, 675-679, DNA sequence gene/EMBL accession N° M63676) gives the following results:

Region from ORF A, SEQ ID NO 4: amino acids 370 - 451: is 50% identical to the acyl carrier protein domain of module 1 of the eryA locus of *Saccharopolyspora erythraea*.
Region from ORF A, SEQ ID NO 4: amino acids 469 - 889: is 65% identical to the ketoacyl synthase domain of module 1 of the eryA locus of *Saccharopolyspora erythraea*.
Region from ORF A, SEQ ID NO 4: amino acids 982 - 1292: is 54% identical to the acyl-transferase domain of module 1 of the eryA locus of *Saccharopolyspora erythraea*.
Region from ORF A, SEQ ID NO 4: amino acids 1324 - 1442: is 42% identical to the dehydratase domain of module 4 of the eryA locus of *Saccharopolyspora erythraea*.
Region from ORF A, SEQ ID NO 4: amino acids 1664 - 1840: is 56% identical to the keto-reductase domain of module 1 of the eryA locus of *Saccharopolyspora erythraea*.
Region from ORF A, SEQ ID NO 4: amino acids 1929 - 2000: is 53% identical to the acyl carrier protein domain of module 1 of the eryA locus of *Saccharopolyspora erythraea*.
Region from ORF A, SEQ ID NO 4: amino acids 2032 - 2453: is 64% identical to the ketoacyl synthase domain of module 1 of the eryA locus of *Saccharopolyspora erythraea*.
Region from ORF A, SEQ ID NO 4: amino acids 2554 - 2965: is 37% identical to the acyl-transferase domain of module 1 of the eryA locus of *Saccharopolyspora erythraea*.
Region from ORF A, SEQ ID NO 4: amino acids 2918 - 2991: is 54% identical to the acyl carrier protein domain of module 1 of the eryA locus of *Saccharopolyspora erythraea*.
Region from ORF A, SEQ ID NO 4: amino acids 3009 - 3431: is 65% identical to the ketoacyl synthase domain of module 1 of the eryA locus of *Saccharopolyspora erythraea*.
Region from ORF A, SEQ ID NO 4: amino acids 3532 - 3847: is 53% identical to the acyl-transferase domain of module 1 of the eryA locus of *Saccharopolyspora erythraea*.
Region of ORF A, SEQ ID NO 4: amino acids 4142 - 4307: is 43% identical to the keto-reductase domain of module 1 of the eryA locus of *Saccharopolyspora erythraea*.
Region of ORF A, SEQ ID NO 4: amino acids 4405 - 4490: is 50% identical to the acyl carrier protein domain of module 1 of the eryA locus of *Saccharopolyspora erythraea*.

In addition to these significant homologies with the eryA polyketide synthase of *S. erythraea*, the region of ORF A, SEQ ID NO 4: amino acids 1 - 356 is 53% identical to the postulated starter unit activation domain of the rapamycin polyketide synthase from *Streptomyces hygroscopicus* (Aparicio et al. GENE (1996) **169**, 9-16)

The great similarities found in the amino acid sequence of the enzymatic domains suggest unambiguously that the protein-encoding region (ORF A) of the *A. mediterranai* chromosomal region depicted in SEQ ID NO 3 codes for a typical modular (type 1) polyketide synthase. This very large *A. mediterranei* polyketide synthase encoded by ORF A comprises three complete bioactive modules which are each responsible for condensation of a C2 unit in the macrolide ring of the molecule and correct modification of the initially formed β-keto groups. Because of the homology with activating domains of the rapamycin polyketide synthase, the first module described above very probably comprises an enzymatic domain for activating the aromatic starter unit of rifamycin biosynthesis, 3-amino-5-hydroxybenzoic acid (Ghisalba et al., Biotechnology of Industrial Antibiotics Vandamme E. J. Ed., Decker Inc. New York, (1984) 281-327).

Example 14: Analysis of a second protein encoding region (ORF B) of the cloned *A. mediterranei* chromosomal region depicted in SEQ ID NO 3

The nucleotide sequence in SEQ ID NO 3 is analysed using the Codonpreference computer program (Genetics Computer Group, University of Wisconsin, 1994). This analysis shows that another large open reading frame (ORF B) which codes for a protein is present in the middle region of the sequence (position 15550 - 30759 including stop codon in SEQ ID NO 3). The codons used in ORF B are typical of actinomycetes genes with a high G+C content.

Comparison of the amino acid sequence of ORF B (SEQ ID NO 5, length 5069 amino acids) with other polyketide synthases and specifically with the very well characterized polyketide synthase of *Saccharopolyspora erythraea* (Donadio, Science, (1991) 252, 675-679, DNA sequence gene/EMBL accession N° M63676) gives the following results:

Region of ORF B, SEQ ID NO 5: amino acids 44 - 468: is 62% identical to the ketoacyl synthase domain of module 1 of the eryA locus of *Saccharopolyspora erythraea*.

Region of ORF B, SEQ ID NO 5: amino acids 571 - 889: is 56% identical to the acyl-transferase domain of module 1 of the eryA locus of *Saccharopolyspora erythraea*.

Region of ORF B, SEQ ID NO 5: amino acids 921 - 1055: is 47% identical to the dehydratase domain of module 4 of the eryA locus of *Saccharopolyspora erythraea*.

Region of ORF B, SEQ ID NO 5: amino acids 1353 - 1525: is 49% identical to the keto-reductase domain of module 1 of the eryA locus of *Saccharopolyspora erythraea*.

Region of ORF B, SEQ ID NO 5: amino acids 1621 - 1706: is 53% identical to the acyl carrier protein domain of module 1 of the eryA locus of *Saccharopolyspora erythraea*.

Region of ORF B, SEQ ID NO 5: amino acids 1726 - 2148: is 62% identical to the ketoacyl synthase domain of module 1 of the eryA locus of *Saccharopolyspora erythraea*.

Region of ORF B, SEQ ID NO 5: amino acids 2251 - 2560: is 55% identical to the acyl-transferase domain of module 1 of the eryA locus of *Saccharopolyspora erythraea*.

Region of ORF B, SEQ ID NO 5: amino acids 2961 - 3132: is 49% identical to the keto-reductase domain of module 1 of the eryA locus of *Saccharopolyspora erythraea*.

Region of ORF B, SEQ ID NO 5: amino acids 3228 - 3313: is 52% identical to the acyl carrier protein domain of module 1 of the eryA locus of *Saccharopolyspora erythraea*.

Region of ORF B, SEQ ID NO 5: amino acids 3332 - 3755: is 63% identical to the ketoacyl synthase domain of module 1 of the eryA locus of *Saccharopolyspora erythraea*.

Region of ORF B, SEQ ID NO 5: amino acids 3857 - 4173: is 52% identical to the acyl-transferase domain of module 1 of the eryA locus of *Saccharopolyspora erythraea*.

Region of ORF B, SEQ ID NO 5: amino acids 4664 - 4799: is 47% identical to the keto-reductase domain of module 1 of the eryA locus of *Saccharopolyspora erythraea*.

Region of ORF B, SEQ ID NO 5: amino acids 4929 - 5014: is 52% identical to the acyl carrier protein domain of module 1 of the eryA locus of *Saccharopolyspora erythraea*.

Example 15: Analysis of a third protein-encoding region (ORF C) of the cloned *A. mediterranei* chromosomal region depicted in SEQ ID NO 3

The nucleotide sequence in SEQ ID NO 3 is analysed using the Codonpreference computer program (Genetics Computer Group, University of Wisconsin, 1994). This analysis shows that a large open reading frame (ORF C) which codes for a protein is present in the middle region of the sequence (position 30895 - 36060 including stop codon in SEQ ID NO 3). The codons used in ORF C are typical of actinomycetes genes with a high G+C content.

Comparison of the amino acid sequence of ORF C (SEQ ID NO 6, length 1721 amino acids) with other polyketide synthases and specifically with the very well characterized polyketide synthase from *Saccharopolyspora erythraea* (Donadio, Science, (1991) 252, 675-679, DNA sequence gene/EMBL accession N° M63676) gives the following results:

Region of ORF C, SEQ ID NO 6: amino acids 1 - 414: is 63% identical to the ketoacyl synthase domain of module 1 of the eryA locus of *Saccharopolyspora erythraea*.
Region of ORF C, SEQ ID NO 6: amino acids 514 - 828: is 54% identical to the acyl-transferase domain of module 1 of the eryA locus of *Saccharopolyspora erythraea*.
Region of ORF C, SEQ ID NO 6: amino acids 1290 - 1399: is 49% identical to the keto-reductase domain of module 1 of the eryA locus of *Saccharopolyspora erythraea*.
Region of ORF C, SEQ ID NO 6: amino acids 1563 - 1648: is 55% identical to the acyl carrier protein domain of module 1 of the eryA locus of *Saccharopolyspora erythraea*.

Example 16: Analysis of a fourth protein-encoding region (ORF D) of the cloned *A. mediterranei* chromosomal region depicted in SEQ ID NO 3

The nucleotide sequence in SEQ ID NO 3 is analysed using the Codonpreference computer program (Genetics Computer Group, University of Wisconsin, 1994). This analysis shows that a large open reading frame (ORF D) which codes for a protein is present in the middle region of the sequence (position 36259 - 41325 including stop codon in SEQ ID NO 3). The codons used in ORF D are typical of actinomycetes genes with a high G+C content.

Comparison of the amino acid sequence of ORF D (SEQ ID NO 7, length 1688 amino acids) with other polyketide synthases and specifically with the very well characterized polyketide synthase from *Saccharopolyspora erythraea* (Donadio, Science, (1991) 252, 675-679, DNA sequence genes/EMBL accession N° M63676) gives the following results:

Region of ORF D, SEQ ID NO 7: amino acids 1 - 418: is 64% identical to the ketoacyl synthase domain of module 1 of the eryA locus of *Saccharopolyspora erythraea*.
Region of ORF D, SEQ ID NO 7: amino acids 524 - 841: is 54% identical to the acyl-transferase domain of module 1 of the eryA locus of *Saccharopolyspora erythraea*.
Region of ORF D, SEQ ID NO 7: amino acids 1260 - 1432: is 51% identical to the keto-reductase domain of module 1 of the eryA locus of *Saccharopolyspora erythraea*.
Region of ORF D, SEQ ID NO 7: amino acids 1523 - 1608: is 53% identical to the acyl carrier protein domain of module 1 of the eryA locus of *Saccharopolyspora erythraea*.

Example 17: Analysis of a fifth protein-encoding region (ORF E) of the cloned *A. mediterranei* chromosomal region depicted in SEQ ID NO 3
The nucleotide sequence in SEQ ID NO 3 is analysed using the Codonpreference computer program (Genetics Computer Group, University of Wisconsin, 1994). This analysis shows that a large open reading frame (ORF E) which codes for a protein is present in the rear region of the sequence (position 41373 - 51614 including stop codon in SEQ ID NO 3). The codons used in ORF E are typical of actinomycetes genes with a high G+C content.

Comparison of the amino acid sequence of ORF E (SEQ ID NO 8, length 3413 amino acids) with other polyketide synthases and specifically with the very well characterized polyketide synthase from *Saccharopolyspora erythraea* (Donadio, Science, (1991) 252, 675-679, DNA sequence gene/EMBL accession N° M63676) gives the following results:

Region of ORF E, SEQ ID NO 8: amino acids 31 - 451: is 64% identical to the ketoacyl synthase domain of module 1 of the eryA locus of *Saccharopolyspora erythraea*.
Region of ORF E, SEQ ID NO 8: amino acids 555 - 874: is 37% identical to the acyl-transferase domain of module 1 of the eryA locus of *Saccharopolyspora erythraea*.
Region of ORF E, SEQ ID NO 8: amino acids 907 - 1036: is 49% identical to the dehydratase domain of module 4 of the eryA locus of *Saccharopolyspora erythraea*.
Region of ORF E, SEQ ID NO 8: amino acids 1336 - 1500: is 52% identical to the keto-reductase domain of module 1 of the eryA locus of *Saccharopolyspora erythraea*.
Region of ORF E, SEQ ID NO 8: amino acids 1598 - 1683: is 51% identical to the acyl carrier protein domain of module 1 of the eryA locus of *Saccharopolyspora erythraea*.
Region of ORF E, SEQ ID NO 8: amino acids 1702 - 2124: is 62% identical to the ketoacyl synthase domain of module 1 of the eryA locus of *Saccharopolyspora erythraea*.

Region of ORF E, SEQ ID NO 8: amino acids 2229 - 2543: is 53% identical to the acyl-transferase domain of module 1 of the eryA locus of *Saccharopolyspora erythraea*.
Region of ORF E, SEQ ID NO 8: amino acids 2573 - 2700: is 47% identical to the dehydratase domain of module 4 of the eryA locus of *Saccharopolyspora erythraea*.
Region of ORF E, SEQ ID NO 8: amino acids 3054 - 3227: is 52% identical to the keto-reductase domain of module 1 of the eryA locus of *Saccharopolyspora erythraea*.
Region of ORF E, SEQ ID NO 8: amino acids 3324 - 3405: is 51% identical to the acyl carrier protein domain of module 1 of the eryA locus of *Saccharopolyspora erythraea*.

Example 18: Analysis of a sixth protein-encoding region (ORF F) of the cloned *A. mediterranei* chromosomal region depicted in SEQ ID NO 3

The nucleotide sequence in SEQ ID NO 3 is analysed using the Codonpreference computer program (Genetics Computer Group, University of Wisconsin, 1994). This analysis shows that an open reading frame (ORF F) which codes for a protein is present in the rear region of the sequence (position 51713 - 52393 including stop codon in SEQ ID NO 3). The codons used in ORF F are typical of actinomycetes genes with a high G+C content.

Comparison of the amino acid sequence of ORF F (SEQ ID NO 9, length 226 amino acids) with proteins from the EMBL databank (Heidelberg) shows a great similarity with the N-hydroxyarylamine O-acyltransferase from *Salmonella typhimurium* (29% identity over a region of 134 amino acids). There is also significant homology with arylamine acyl-transferases from other organisms. It can be concluded from these agreements that the ORF F found in *A. mediterranei* in SEQ ID No 3 codes for an arylamine acyl transferase, and it can be assumed that this enzyme is responsible for the linkage of the long acyl chain produced by the polyketide synthase to the amino group on the starter molecule, 3-amino-5-hydroxybenzoic acid. This reaction would close the rifamycin ring system correctly after completion of the condensation steps by the polyketide synthase.

Example 19: Summarizing assessment of the function of the proteins encoded by ORF A - F in SEQ ID NO 3, and their role in the biosynthesis of rifamycin

The five protein-encoding regions (ORF A-E), described in Examples 13 - 17, of SEQ ID NO 3 comprise proteins with very great similarity (in the amino acid sequence and the arrangement of the enzymatic domains) to polyketide synthases for polyketides of the macrolide type. Taken together, these five multifunctional enzymes comprise 10 polyketide

synthase modules which are each responsible for a condensation step in the polyketide synthesis. 10 such condensation steps are likewise necessary for rifamycin biosynthesis (Ghisalba et al., Biotechnology of Industrial Antibiotics Vandamme E. J. Ed., Decker Inc. New York, (1984) 281-327). The processing of the particular keto groups required by the enzymatic domains within the modules substantially corresponds to the activity required by the rifamycin molecule, if it is assumed that the polyketide synthesis takes place "colinearly" with the arrangement of the modules in the gene cluster of *A. mediterranei* (this is so for other macrolide antibiotics such as erythromycin and rapamycin). It may be added here that it is not certain whether transcription of the five ORFs results in five proteins; in particular, ORF C and ORF D might possibly be translated to a large protein.

An enzymatic domain which is very probably responsible for activating the starter molecule, 3-hydroxy-5-aminobenzoic acid, of rifamycin biosynthesis can be found at the N terminus of ORF A, the start of the polyketide synthase. Directly below the described rifamycin polyketide synthase gene cluster there is a gene (ORF F) which very probably determines a protein which brings about ring closure of the rifamycin molecule after completion of the condensation steps by the polyketide synthase.

It can be concluded on the basis of these findings that the *A. mediterranei* chromosomal region described in SEQ ID NO 3 is responsible for the ten condensation steps required for rifamycin polyketide synthesis, including activation of the starter molecule 3-hydroxy-5-aminobenzoic acid, and the concluding ring closure.

Deposited microorganisms

The following microorganisms and plasmids have been deposited at the Deutsche Sammlung von Mikroorganismen und Zellkulturen GmbH (DSM), Mascheroder Weg 1b, D-38124 Braunschweig, in accordance with the requirements of the Budapest Treaty.

Microorganism/Plasmid	Date of deposit	Deposit number
E. coli with plasmid pRi7-3	10.08.96	DSM 11114
E. coli with plasmid pNE112	14.07.97	DSM 11657
E. coli with plasmid pNE95	14.07.97	DSM 11656
E. coli with plasmid pRi44-2	14.07.97	DSM 11655

SEQUENCE LISTING

(1) GENERAL INFORMATION:

(i) APPLICANT:
(A) NAME: Novartis AG
(B) STREET: Schwarzwaldallee 215
(C) CITY: Basel
(E) COUNTRY: Switzerland
(F) POSTAL CODE (ZIP): 4058
(G) TELEPHONE: +41 61 324 1111
(H) TELEFAX: + 41 61 322 75 32

(ii) TITLE OF INVENTION: Rifamycin biosynthesis gene cluster

(iii) NUMBER OF SEQUENCES: 9

(iv) COMPUTER READABLE FORM:
(A) MEDIUM TYPE: Floppy disk
(B) COMPUTER: IBM PC compatible
(C) OPERATING SYSTEM: PC-DOS/MS-DOS
(D) SOFTWARE: PatentIn Release #1.0, Version #1.30 (EPO)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 1:

GGTACCCGGT GTTCGCGACG GCGTTCGACG AGGCTTGCGA GCAGCTGGAC GTCTGTCTGG 60

CCGGCCGTGC CGGGCACCGC GTGCGGGACG TCGTGCTCGG CGAAGTGCCC GCCGAAACCG 120

GGCTGCTGAA CCAGACGGTC TTCACCCAAG CCGGGCTGTT CGCGGTGGAG AGCGCGCTGT 180

TCCGGCTCGC CGAATCCTGG GGTGTCCGGC CGGACGTGGT GCTCGGCCAC TCCATCGGGG 240

AGATCACCGC CGCGTATGCC GCGGGCGTCT TCTCGCTGCC GGACGCCGCC CGGATCGTCG 300

CGGCGCGCGG CCGGCTGATG CAGGCGCTGG CGCCGGGCGG GGCGATGGTC GCCGTCGCCG 360

CCTCCGAAGC CGAGGTGGCC GAACTGCTCG GCGACGGCGT GGAACTCGCC GCCGTCAACG 420

GCCCTTCGGC GGTAGTCCTT TCCGGGGACG CGGACGCGGT CGTCGCGGCC GCCGCCCGCA 480

TGCGCGAGCG CGGGCACAAG ACCAAGCAGC TCAAGGTTTC GCACGCGTTC CACTCCGCGC 540

GGATGGCGCC GATGCTGGCG GAGTTCGCCG CCGAGCTGGC CGGCGTGACG TGGCGCGAGC 600

CGGAGATCCC GGTGGTCTCC AACGTGACCG GCCGGTTCGC CGAGCCCGGC GAACTGACCG 660

AGCCGGGCTA CTGGCCGAG CACGTGCGGC GCCGGTGCG GTTCGCCGAG GGCGTCGCGG 720

CGGCGACGGA GTCCGGCGGC TCGCTGTTCG TGGAGCTCGG GCCGGGGGCG GCGCTGACCG 780

CCCTCGTCGA GGAGACGGCC GAGGTCACCT GCGTCGCGGC CCTGCGGGAC GACCGCCCGG 840

AGGTCACCGC GCTGATCACC GCGGTCGCCG AGCTGTTCGT CCGCGGGGTT GCGGTCGATT 900

GCCCGGCCCT GCTGCCGCCG GTCACCGGGT TCGTCGACCT GCCGAAGTAC GCCTTCGACC 960

AGCAGCACTA TTGGCTGCAG CCCGCCGCGC AGGCCACGGA CGCGGCCTCG CTCGGGCAGG 1020

TCGCGGCCGA CCACCCGCTG CTGGGCGCGG TGGTCCGGCT GCCGCAGTCG GACGGCCTGG 1080

TCTTCACCTC GCGGCTGTCA TTGAAATCGC ACCCGTGGCT GGCCGACCAC GTCATCGGCG 1140

GGGTGGTGCT CGTCGCGGGC ACCGGGCTCG TCGAGCTGGC CGTCCGGGCC GGGGACGAGG 1200

CCGGCTGCCC GGTCCTCGAA GAACTCGTCA TCGAGGCTCC GCTGGTCGTC CCCGACCACG 1260

GCGGGGTCCG GATCCAGGTC GTCGTGGGGG CACCGGGGGA GACCGGTTCG CGCGCGGTCG 1320

AGGTGTACTC CCTGCCCGAG GACGCCGGTG CCGAAGTCGTG GCCCCGGCAC GCCACCGGGT 1380

TCCTGGCTGC GACGCCGTCG CAGCACAAGC CGTTCGACTT CACCGCCTGG CCGCCGCCCG 1440

GCGTCGAGCG CGTCGACGTC GAGGACTTCT ACGACGGCTT CGTCGACCGC GGGTACGCCT 1500

ACGGGCCGTC GTTCCGGGGC CTGCGGGCGG TGTGGCGGCG CGGCGACGAA GTGTTCGCCG 1560

AGGTCGCCCT GGCCGAGGAC GACCGCGCGG ACGCGGCCCG GTTCGGCATC CACCCCGGCC 1620

TGCTGGACGC CGCCCTGCAC GCGGGCATGG CCGGTGCCAC CACCACGGAA GAGCCCGGCC 1680

GGCCGGTGCT GCCGTTCGCC TGGAACGGCC TGGTGCTGCA CGCGGCCGGG GCGTCCGCGC 1740

TGCGGGTCCG GCTCGCCCCG AGCGGTCCGG ACGCCCTGTC GGTCGAGGCC GCGGACGAGG 1800

CCGGCGGTCT CGTTGTGACG GCGGACTCGC TGGTCTCCCG GCCGGTGTCG GCCGAACAGC 1860

TGGGCGCGGC GCCGAACCAC GACGCGTTGT TCCGCGTGGA GTGGACCGAG ATTTCCTCGG 1920

CTGGAGACGT TCCGGCGGAC CACGTCGAAG TGCTCGAAGC CGTCGGCGAG GATCCCCTGG 1980

AACTGACCGG CCGGGTCCTG GAGGCCGTGC AGACCTGGCT CGCCGACGCA GCCGACGACG 2040

CTCGCCTGGT CGTGGTGACC CGCGGCGCCG TCCACGAGGT GACTGACCCG GCCGGTGCCG 2100

CGGTGTGGGG CCTGATCCGG GCCGCGCAGG CGGAAAACCC GGACCGGATC GTGCTGCTGG 2160

ACACCGACGG TGAAGTGCCG CTAGGCCGGG TGCTGGCCAC CGGCGAGCCC CAAACAGCCG 2220

TCCGAGGCGC CACGCTGTTC GCCCCGCGGC TGGCCCGCGC CGAGGCCGCG GAGGCACCGG 2280

CAGTGACCGG CGGGACGGTC CTGATCTCGG GCGCCGGCTC GCTGGGCGCG CTCACCGCCC 2340

GGCACCTGGT CGCCCGGCAC GGAGTCCGGC GGCTGGTGCT CGTCAGCCGC CGTGGCCCCG 2400

ACGCCGACGG CATGGCCGAA CTGACCGCTG AACTCATCGC TCAGGGCGCC GAGGTCGCCG 2460

TAGTCGCTTG CGACCTGGCC GACCGGGACC AGGTCCGGGT ACTGCTGGCC GAGCACCGCC 2520

CGAACGCCGT CGTGCACACG GCCGGTGTTC TCGACGACGG CGTCTTCGAG TCGCTGACGC 2580

GGGAGCGGCT GGCCAAGGTC TTCGCGCCCA AAGTTACTGC TGCCAATCAC CTCGACGAGC 2640

TGACCCGCGA ACTGGATCTT CGCGCGTTCG TCGTGTTCTC CTCCGCCTCC GGGGTCTTCG 2700

GCTCCGCCGG GCAGGGCAAC TACGCCGCTG CCAACGCCTA CCTGGACGCC GTGGTCGCCA 2760

ACCGCCGGGC CGCGGGCCTG CCCGGCACAT CGCTGGCCTG GGGCCTGTGG GAACAGACCG 2820

ACGGGATGAC CGCGCACCTC GGCGACGCCG ACCAGGCGCG GGCGAGTCGC GGCGGGGTCC 2880

TGGCCATCTC ACCCGCCGAA GGCATGGAGC TGTTCGACGC AGCGCCGGAC GGGCTCGTCG 2940

TCCCGGTCAA GCTGGACCTG CGCAAGACCC GCGCCGGCGG GACGGTGCCG CACCTGCTGC 3000

GCGGCCTGGT CCGCCCGGGA CGGCAGCAGG CCCGTCCGGC GTCCACTGTG GACAACGGAC 3060

TGGCCGGGCG ACTCGCCGGG CTCGCGCCGG CGGAGCAGGA GGCGCTGCTG CTCGACGTCG 3120

TCCGCACGCA GGTCGCGCTG GTGCTCGGGC ACGCCGGGCC GGAGGCCGTC CGCGCGGACA 3180

CGGCGTTCAA GGACACCGGC TTCGACTCGC TGACGTCGGT GGAACTGCGC AACCGGCTGC 3240

GCGAGGCGAG CGGGCTGAAG CTGCCCGCGA CGCTCGTCTT CGACTACCCG ACGCCGGTCG 3300

CGCTGGCCCG CTACCTGCGT GACGAATTCG GCGACACGGT GGCAACAACT CCGGTGGCCA 3360

CCGCGGCCGC AGCGGACGCC GGCGAGCCGA TCGCCATCGT CGGCATGGCG TGCCGGCTGC 3420

CGGGCGGGGT CACCGATCCC GAAGGCCTGT GGCGCCTGGT GCGCGACGGC CTCGAAGGGC 3480

TGTCTCCCTT CCCCGAGGAC CGGGGCTGGG ACCTGGAGAA CCTGTTCGAC GACGACCCCG 3540

ACCGCTCCGG CACGACGTAC ACCAGCCGGG GCGGGTTCCT CGACGGCGCC GGCCTGTTCG 3600

ACGCGGGCTT CTTCGGATT TCGCCGCGCG AGGCGCTGGC CATGGACCCG CAGCAGCGGC 3660

TGCTGCTCGA GCCGGCCTGG GAAGCCCTCG AAGGCACCGG TGTCGACCCG GGCTCGTTGA 3720

AGGGCGCCGA CGTCGGGGTG TTCGCCGGGG TGTCCAACCA GGGCTATGGG ATGGGCGCGG 3780

ATCCGGCCGA ACTGGCGGGG TACGCGAGCA CGGCGGGCGC TTCGAGCGTC GTCTCGGGCC 3840

GAGTCTCGTA CGTCTTCGGG TTCGAAGGAC CGGCGGTCAC GATCGACACG GCTTGCTCGT 3900

CGTCGCTGGT GCCGATGCAC CTGGCCGGGC AGGCGCTGCG GCAGGGCGAG TGCTCGATGG 3960

CCCTGGCCGG TGGCGTCACG GTGATGGGGA CGCCCGGCAC GTTCGTGGAG TTCGCGAAGC 4020

AGCGCGGCCT GGCCGGCGAC GGCCGGTGCA AGCCCTACGC CGAAGGCGCG GACGGCACGG 4080

GCTGGGCCGA GGCCGTCGGG GTCGTCGTGC TGGAGCGGCT GTCGGTGGCG CGCGAGCGCG 4140

GGCACCGGGT GCTGGCCGTG CTGCGCGGCA GCGCGGTCAA CTCCGACGGC GCGTCCAACG 4200

GCCTGACCGC CCCCAACGGG CCGTCGCAGC AACGGGTGAT CCGCCGGGCC CTGGCCGCCG 4260

CCGGCCTCGA ACCGTCCGAT GTGGACATCG TGGAAGGGCA CGGCACCGGC ACGGCGCTGG 4320

GCGACCCGAT CGAGGCGCAG GCCCTGCTGG CCACCTACGG CAAGGACCGC GACCCGGAGA 4380

CGCCGTTGTG GCTGGGGTCG GTGAAGTCGA ACTTCGGCCA CACGCAGTCC GCGGCCGGCG 4440

TGGCCGGGGT GATCAAGATG GTGCAGGCGC TGCGCCACGG CGTCATGCCG CCCACCCTGC 4500

ACGTGGACCG GCCCACCAGC CAGGTCGACT GGTCCGCGGG GGCCGTCGAA GTGCTGACCG 4560

AGGCACGGGA GTGGCCGCGG AACGGCCGTC CGCGCCGGGC CGGGGTGTCC TCGTTCGGGA 4620

TCAGCGGCAC GAACGCCCAC CTGATCATCG AAGAAGCACC GGCCGAGCCA CAGCTTGCCG 4680

GACCACCGCC GGACGGCGGT GTGGTGCCGC TGGTCGTCTC GGCTCGCAGC CCCGGTGCCC 4740

TGGCCGGTCA GGCGCGTCGG CTGGCCACGT TCCTCGGCGA CGGCCCCCTT TCCGACGTCC 4800

CCGGTGCGCT GACGAGCCGC GCCCTGTTCG GCGAGCGCGC GGTCGTCGTG GCGGATCGG 4860

CCGAGGAAGC CCGCGCCGGT CTGGGCGCAC TGGCCCGCGG CGAAGACGCG CCGGGCCTGG 4920

TCCGCGGCCG GGTGCCCGCG TCCGGCCTGC CGGCAAGCT CGTGTGGGTG TTCCCCGGGC 4980

AGGGGACGCA GTGGGTGGGC ATGGGCCGCG AACTCCTCGA AGAGTCTCCG GTGTTCGCCG 5040

AGCGGATCGC CGAGTGTGCG GCCGCGCTGG AGCCGTGGAT CGGCTGGTCG CTGTTCGACG 5100

TCCTCCGTGG CGACGGTGAC CTCGATCGGG TCGATGTGCT GCAGCCCGCG TGCTTTGCGG 5160

TGATGGTCGG CTTGGCCGCG GTGTGGTCCT CGGCCGGGGT GGTCCCCGAT GCGGTGCTCG 5220

GCCACTCCCA GGGTGAGATC GCCGCGGGGT GCGTGTCGGG TGCGTTCTCG CTGGAGGATG 5280

CGGCGAAGGT GGTTGCCCTG CGCAGCCAGG CCATCGCCGC GAAGCTCTCC GGCCGCGGCG 5340

GGATGGCTTC GGTCGCCTTG GGCGAAGCCG ATGTGGTGTC GCGGCTGGCG GACGGGGTCG 5400

AGGTGGCTGC CGTCAACGGT CCGGCGTCCG TGGTGATCGC GGGGGATGCC CAGGCCCTCG 5460

ACGAAACGCT GGAAGCGCTG TCCGGTGCGG GAATCCGGGC TCGGCGGGTG GCGGTGGACT 5520

ACGCCTCGCA CACCCGGCAC GTCGAAGACA TCGAAGACAC CCTCGCCGAA GCGCTGGCCG 5580

GGATCGACGC CCGGGCGCCG CTGGTGCCGT TCCTCICCAC CCTCACCGGC GAGTGGATCC 5640

GGGACGAGGG CGTCGTGGAC GGCGGCTACT GGTACC 5676

(2) INFORMATION FOR SEQ ID NO: 2:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 1891 amino acids
(B) TYPE: amino acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 2:

Tyr Pro Val Phe Ala Thr Ala Phe Asp Glu Ala Cys Glu Gln Leu Asp
1 5 10 15

Val Cys Leu Ala Gly Arg Ala Gly His Arg Val Arg Asp Val Val Leu
20 25 30

Gly Glu Val Pro Ala Glu Thr Gly Leu Leu Asn Gln Thr Val Phe Thr
35 40 45

Gln Ala Gly Leu Phe Ala Val Glu Ser Ala Leu Phe Arg Leu Ala Glu
50 55 60

Ser Trp Gly Val Arg Pro Asp Val Val Leu Gly His Ser Ile Gly Glu
65 70 75 80

Ile Thr Ala Ala Tyr Ala Ala Gly Val Phe Ser Leu Pro Asp Ala Ala
85 90 95

Arg Ile Val Ala Ala Arg Gly Arg Leu Met Gln Ala Leu Ala Pro Gly
100 105 110

Gly Ala Met Val Ala Val Ala Ala Ser Glu Ala Glu Val Ala Glu Leu
115 120 125

Leu Gly Asp Gly Val Glu Leu Ala Ala Val Asn Gly Pro Ser Ala Val
130 135 140

Val Leu Ser Gly Asp Ala Asp Ala Val Val Ala Ala Ala Ala Arg Met
145 150 155 160

Arg Glu Arg Gly His Lys Thr Lys Gln Leu Lys Val Ser His Ala Phe
165 170 175

His Ser Ala Arg Met Ala Pro Met Leu Ala Glu Phe Ala Ala Glu Leu
180 185 190

Ala Gly Val Thr Trp Arg Glu Pro Glu Ile Pro Val Val Ser Asn Val
195 200 205

Thr Gly Arg Phe Ala Glu Pro Gly Glu Leu Thr Glu Pro Gly Tyr Trp
210 215 220

Ala Glu His Val Arg Arg Pro Val Arg Phe Ala Glu Gly Val Ala Ala
225 230 235 240

Ala Thr Glu Ser Gly Gly Ser Leu Phe Val Glu Leu Gly Pro Gly Ala
245 250 255

Ala Leu Thr Ala Leu Val Glu Glu Thr Ala Glu Val Thr Cys Val Ala
260 265 270

Ala Leu Arg Asp Asp Arg Pro Glu Val Thr Ala Leu Ile Thr Ala Val
275 280 285

Ala Glu Leu Phe Val Arg Gly Val Ala Val Asp Trp Pro Ala Leu Leu
290 295 300

Pro Pro Val Thr Gly Phe Val Asp Leu Pro Lys Tyr Ala Phe Asp Gln
305 310 315 320

Gln His Tyr Trp Leu Gln Pro Ala Ala Gln Ala Thr Asp Ala Ala Ser
325 330 335

Leu Gly Gln Val Ala Ala Asp His Pro Leu Leu Gly Ala Val Val Arg
340 345 350

Leu Pro Gln Ser Asp Gly Leu Val Phe Thr Ser Arg Leu Ser Leu Lys
355 360 365

Ser His Pro Trp Leu Ala Asp His Val Ile Gly Gly Val Val Leu Val
370 375 380

Ala Gly Thr Gly Leu Val Glu Leu Ala Val Arg Ala Gly Asp Glu Ala
385 390 395 400

Gly Cys Pro Val Leu Glu Glu Leu Val Ile Glu Ala Pro Leu Val Val
405 410 415

Pro Asp His Gly Gly Val Arg Ile Gln Val Val Val Gly Ala Pro Gly
420 425 430

Glu Thr Gly Ser Arg Ala Val Glu Val Tyr Ser Leu Arg Glu Asp Ala
435 440 445

Gly Ala Glu Val Trp Ala Arg His Ala Thr Gly Phe Leu Ala Ala Thr
450 455 460

Pro Ser Gln His Lys Pro Phe Asp Phe Thr Ala Trp Pro Pro Pro Gly

465 470 475 480

Val Glu Arg Val Asp Val Glu Asp Phe Tyr Asp Gly Phe Val Asp Arg
485 490 495

Gly Tyr Ala Tyr Gly Pro Ser Phe Arg Gly Leu Arg Ala Val Trp Arg
500 505 510

Arg Gly Asp Glu Val Phe Ala Glu Val Ala Leu Ala Glu Asp Asp Arg
515 520 525

Ala Asp Ala Ala Arg Phe Gly Ile His Pro Gly Leu Leu Asp Ala Ala
530 535 540

Leu His Ala Gly Met Ala Gly Ala Thr Thr Thr Glu Glu Pro Gly Arg
545 550 555 560

Pro Val Leu Pro Phe Ala Trp Asn Gly Leu Val Leu His Ala Ala Gly
565 570 575

Ala Ser Ala Leu Arg Val Arg Leu Ala Pro Ser Gly Pro Asp Ala Leu
580 585 590

Ser Val Glu Ala Ala Asp Glu Ala Gly Gly Leu Val Val Thr Ala Asp
595 600 605

Ser Leu Val Ser Arg Pro Val Ser Ala Glu Gln Leu Gly Ala Ala Ala
610 615 620

Asn His Asp Ala Leu Phe Arg Val Glu Trp Thr Glu Ile Ser Ser Ala
625 630 635 640

Gly Asp Val Pro Ala Asp His Val Glu Val Leu Glu Ala Val Gly Glu
645 650 655

Asp Pro Leu Glu Leu Thr Gly Arg Val Leu Glu Ala Val Gln Thr Trp
660 665 670

Leu Ala Asp Ala Ala Asp Asp Ala Arg Leu Val Val Val Thr Arg Gly
675 680 685

Ala Val His Glu Val Thr Asp Pro Ala Gly Ala Ala Val Trp Gly Leu
690 695 700

Ile Arg Ala Ala Gln Ala Glu Asn Pro Asp Arg Ile Val Leu Leu Asp
705 710 715 720

Thr Asp Gly Glu Val Pro Leu Gly Arg Val Leu Ala Thr Gly Glu Pro
725 730 735

Gln Thr Ala Val Arg Gly Ala Thr Leu Phe Ala Pro Arg Leu Ala Arg
740 745 750

Ala Glu Ala Ala Glu Ala Pro Ala Val Thr Gly Gly Thr Val Leu Ile
755 760 765

Ser Gly Ala Gly Ser Leu Gly Ala Leu Thr Ala Arg His Leu Val Ala
770 775 780

Arg His Gly Val Arg Arg Leu Val Leu Val Ser Arg Arg Gly Pro Asp
785 790 795 800

Ala Asp Gly Met Ala Glu Leu Thr Ala Glu Leu Ile Ala Gln Gly Ala
805 810 815

Glu Val Ala Val Val Ala Cys Asp Leu Ala Asp Arg Asp Gln Val Arg
820 825 830

Val Leu Leu Ala Glu His Arg Pro Asn Ala Val Val His Thr Ala Gly
835 840 845

Val Leu Asp Asp Gly Val Phe Glu Ser Leu Thr Arg Glu Arg Leu Ala
850 855 860

```
Lys Val Phe Ala Pro Lys Val Thr Ala Ala Asn His Leu Asp Glu Leu
865                 870                 875                 880

Thr Arg Glu Leu Asp Leu Arg Ala Phe Val Val Phe Ser Ser Ala Ser
                885                 890                 895

Gly Val Phe Gly Ser Ala Gly Gln Gly Asn Tyr Ala Ala Ala Asn Ala
            900                 905                 910

Tyr Leu Asp Ala Val Val Ala Asn Arg Arg Ala Ala Gly Leu Pro Gly
        915                 920                 925

Thr Ser Leu Ala Trp Gly Leu Trp Glu Gln Thr Asp Gly Met Thr Ala
    930                 935                 940

His Leu Gly Asp Ala Asp Gln Ala Arg Ala Ser Arg Gly Gly Val Leu
945                 950                 955                 960

Ala Ile Ser Pro Ala Glu Gly Met Glu Leu Phe Asp Ala Ala Pro Asp
                965                 970                 975

Gly Leu Val Val Pro Val Lys Leu Asp Leu Arg Lys Thr Arg Ala Gly
            980                 985                 990

Gly Thr Val Pro His Leu Leu Arg Gly Leu Val Arg Pro Gly Arg Gln
        995                 1000                1005

Gln Ala Arg Pro Ala Ser Thr Val Asp Asn Gly Leu Ala Gly Arg Leu
    1010                1015                1020

Ala Gly Leu Ala Pro Ala Glu Gln Glu Ala Leu Leu Leu Asp Val Val
1025                1030                1035                1040

Arg Thr Gln Val Ala Leu Val Leu Gly His Ala Gly Pro Glu Ala Val
                1045                1050                1055

Arg Ala Asp Thr Ala Phe Lys Asp Thr Gly Phe Asp Ser Leu Thr Ser
```

1060 1065 1070

Val Glu Leu Arg Asn Arg Leu Arg Glu Ala Ser Gly Leu Lys Leu Pro
1075 1080 1085

Ala Thr Leu Val Phe Asp Tyr Pro Thr Pro Val Ala Leu Ala Arg Tyr
1090 1095 1100

Leu Arg Asp Glu Phe Gly Asp Thr Val Ala Thr Thr Pro Val Ala Thr
1105 1110 1115 1120

Ala Ala Ala Ala Asp Ala Gly Glu Pro Ile Ala Ile Val Gly Met Ala
1125 1130 1135

Cys Arg Leu Pro Gly Gly Val Thr Asp Pro Glu Gly Leu Trp Arg Leu
1140 1145 1150

Val Arg Asp Gly Leu Glu Gly Leu Ser Pro Phe Pro Glu Asp Arg Gly
1155 1160 1165

Trp Asp Leu Glu Asn Leu Phe Asp Asp Asp Pro Asp Arg Ser Gly Thr
1170 1175 1180

Thr Tyr Thr Ser Arg Gly Gly Phe Leu Asp Gly Ala Gly Leu Phe Asp
1185 1190 1195 1200

Ala Gly Phe Phe Gly Ile Ser Pro Arg Glu Ala Leu Ala Met Asp Pro
1205 1210 1215

Gln Gln Arg Leu Leu Leu Glu Ala Ala Trp Glu Ala Leu Glu Gly Thr
1220 1225 1230

Gly Val Asp Pro Gly Ser Leu Lys Gly Ala Asp Val Gly Val Phe Ala
1235 1240 1245

Gly Val Ser Asn Gln Gly Tyr Gly Met Gly Ala Asp Pro Ala Glu Leu
1250 1255 1260

```
Ala Gly Tyr Ala Ser Thr Ala Gly Ala Ser Ser Val Val Ser Gly Arg
1265                1270                1275                1280

Val Ser Tyr Val Phe Gly Phe Glu Gly Pro Ala Val Thr Ile Asp Thr
                1285                1290                1295

Ala Cys Ser Ser Ser Leu Val Ala Met His Leu Ala Gly Gln Ala Leu
            1300                1305                1310

Arg Gln Gly Glu Cys Ser Met Ala Leu Ala Gly Gly Val Thr Val Met
        1315                1320                1325

Gly Thr Pro Gly Thr Phe Val Glu Phe Ala Lys Gln Arg Gly Leu Ala
    1330                1335                1340

Gly Asp Gly Arg Cys Lys Ala Tyr Ala Glu Gly Ala Asp Gly Thr Gly
1345                1350                1355                1360

Trp Ala Glu Gly Val Gly Val Val Val Leu Glu Arg Leu Ser Val Ala
                1365                1370                1375

Arg Glu Arg Gly His Arg Val Leu Ala Val Leu Arg Gly Ser Ala Val
            1380                1385                1390

Asn Ser Asp Gly Ala Ser Asn Gly Leu Thr Ala Pro Asn Gly Pro Ser
        1395                1400                1405

Gln Gln Arg Val Ile Arg Arg Ala Leu Ala Gly Ala Gly Leu Glu Pro
    1410                1415                1420

Ser Asp Val Asp Ile Val Glu Gly His Gly Thr Gly Thr Ala Leu Gly
1425                1430                1435                1440

Asp Pro Ile Glu Ala Gln Ala Leu Leu Ala Thr Tyr Gly Lys Asp Arg
                1445                1450                1455
```

Asp Pro Glu Thr Pro Leu Trp Leu Gly Ser Val Lys Ser Asn Phe Gly
1460 1465 1470

His Thr Gln Ser Ala Ala Gly Val Ala Gly Val Ile Lys Met Val Gln
1475 1480 1485

Ala Leu Arg His Gly Val Met Pro Pro Thr Leu His Val Asp Arg Pro
1490 1495 1500

Thr Ser Gln Val Asp Trp Ser Ala Gly Ala Val Glu Val Leu Thr Glu
1505 1510 1515 1520

Ala Arg Glu Trp Pro Arg Asn Gly Arg Pro Arg Arg Ala Gly Val Ser
1525 1530 1535

Ser Phe Gly Ile Ser Gly Thr Asn Ala His Leu Ile Ile Glu Glu Ala
1540 1545 1550

Pro Ala Glu Pro Gln Leu Ala Gly Pro Pro Pro Asp Gly Gly Val Val
1555 1560 1565

Pro Leu Val Val Ser Ala Arg Ser Pro Gly Ala Leu Ala Gly Gln Ala
1570 1575 1580

Arg Arg Leu Ala Thr Phe Leu Gly Asp Gly Pro Leu Ser Asp Val Ala
1585 1590 1595 1600

Gly Ala Leu Thr Ser Arg Ala Leu Phe Gly Glu Arg Ala Val Val Val
1605 1610 1615

Ala Asp Ser Ala Glu Glu Ala Arg Ala Gly Leu Gly Ala Leu Ala Arg
1620 1625 1630

Gly Glu Asp Ala Pro Gly Leu Val Arg Gly Arg Val Pro Ala Ser Gly
1635 1640 1645

Leu Pro Gly Lys Leu Val Trp Val Phe Pro Gly Gln Gly Thr Gln Trp

1650 1655 1660

Val Gly Met Gly Arg Glu Leu Leu Glu Glu Ser Pro Val Phe Ala Glu
1665 1670 1675 1680

Arg Ile Ala Glu Cys Ala Ala Ala Leu Glu Pro Trp Ile Gly Trp Ser
1685 1690 1695

Leu Phe Asp Val Leu Arg Gly Asp Gly Asp Leu Asp Arg Val Asp Val
1700 1705 1710

Leu Gln Pro Ala Cys Phe Ala Val Met Val Gly Leu Ala Ala Val Trp
1715 1720 1725

Ser Ser Ala Gly Val Val Pro Asp Ala Val Leu Gly His Ser Gln Gly
1730 1735 1740

Glu Ile Ala Ala Ala Cys Val Ser Gly Ala Leu Ser Leu Glu Asp Ala
1745 1750 1755 1760

Ala Lys Val Val Ala Leu Arg Ser Gln Ala Ile Ala Ala Lys Leu Ser
1765 1770 1775

Gly Arg Gly Gly Met Ala Ser Val Ala Leu Gly Glu Ala Asp Val Val
1780 1785 1790

Ser Arg Leu Ala Asp Gly Val Glu Val Ala Ala Val Asn Gly Pro Ala
1795 1800 1805

Ser Val Val Ile Ala Gly Asp Ala Gln Ala Leu Asp Glu Thr Leu Glu
1810 1815 1820

Ala Leu Ser Gly Ala Gly Ile Arg Ala Arg Arg Val Ala Val Asp Tyr
1825 1830 1835 1840

Ala Ser His Thr Arg His Val Glu Asp Ile Glu Asp Thr Leu Ala Glu
1845 1850 1855

Ala Leu Ala Gly Ile Asp Ala Arg Ala Pro Leu Val Pro Phe Leu Ser
1860 1865 1870

Thr Leu Thr Gly Glu Trp Ile Arg Asp Glu Gly Val Val Asp Gly Gly
1875 1880 1885

Tyr Trp Tyr
1890

(2) INFORMATION FOR SEQ ID NO: 3:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 53789 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: double
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 3:

GAATTCCAGG CCGTCGACGG CTGCGACATC GCGGTCTTCC GGTGGTCGCA CCGCACGAAG 60

ATCGCCGAAT AAGAATTTCC GGATCTCCCA CGGGAAAGGT TTCCATGACC GACGCAATAT 120

CCTTCGAGGT GCCGTGGGAC CGGACCGACA AGTTCGACCC GCCCGCGGTG TTCGACTCTC 180

TGCGCGAAGA ACGTCCGCTC GCGAAGATGG TTTACCCGGA TGGGCACGTC GGCTGGATCG 240

TTTCCAGCTA CGAGCTGGTC CGCGAGGTCC TCAGCGACCT GCGGTTCAGC CACAGCTGCG 300

AAGTCGGCCA CTTCCCGGTG ACCCACCAGG GCCAGGTCAT CCCGACCCAC CGCCTGATCC 360

CCGGCATGTT CATCCACATG GACCCGCCCG AGCACACGCG CTACGGCAAG CTGCTGACCG 420

GCGAGTTCAC CGTCCGCCGC GCCAGCAGGC TGATCCCGCG GGCCGAGGCC GTGGCCGCCG 480

AGCAGATCGA GGTCATGCGG GCCAAGGGCG CCCCCGCGGA CGTGGTCATG GACTTCGCCA 540

AGCCGCTGGT GCTGCCGATG CTGGGCGAGC TCGTCGGCCT GCCCTACGAG GAACGCGACC 600

GGTACGTGCC CGCGGTGACC CTCCTGCACG ACGCCGAAGC GGACCCGGCC GAGGCCGCGG 660

CCGCCTACGA GGTGGCCGGG AAGTTCTTCG ACGAGGTCAT CGAGCGCCGC CGGCAGCGGC 720

CCCAGGACGA CCTCATCAGC TCGCTCGTCA CCGAGGACCT GAGCCAGGAG GAGCTGCGCA 780

ACATCGTCAC CCTGCTGCTG TTCGCCGGGT ACGAGACCAC CGAGGGCGCG CTCGCCACCG 840

GCGTCTTCGC GCTGCTGCAC CACACCGATC AGCTGGCGGC ACTGCGCGCG GAGCCGGAAA 900

AGCTCGACGC CGCGATCGAA GAGCTGCTGC GCTACCTGAC CGTCAACCAG TACCACACCT 960

ACCGCACCGC GCTGGAGGAC GTGAAGCTGG AGGGCGAGCT GATCAAGAAG GGCGACACGG 1020

TGACGGTGTC GCTGCCCGCG GCCAACCGCG ACCCGGCCAA GTTCGGCTGT CCCGCGGAGC 1080

TCGACATCGA GCGGGACACC TCCGGCCACG TCGCGTTCGG CTTCGGCATC CACCAGTGCC 1140

TGGGCCAGAA CCTGGCGCGC ATCGAGCTGC GGGCCGGCTT CACGGCGCTC CTGCGGGCGT 1200

TCCCCGAGCT CCGGCTGGCC GTCCCGGCCG ACGAGGTTCC GCTGCGGCTG AAGGGTTCCG 1260

TCTTCTCGGT GAAGAAGCTG CCCGTCTCCT GGTGAGCGTT CTTCCCCTCG AACACCCGAA 1320

AGGATCTGCG GCACAGTGCG CACCGATCTC ATCAAGCCAC TTCACGTCGC ACTCCTGGAG 1380

AACGCGACCC GCTTCGCCGG CAAGCCGGCC TTCGCCGACG ACCACCGGAC GGTCACCTAC 1440

GGCGACCTCG AGGCGCGGAC GCGCCGGCTG GCCGGGCACC TGCCGGCCT CGGTGTCCGG 1500

CACGGCGACC GGGTGGCGAT CTGCCTCGGC AACCGGGTGT CCACTGTGGA GAGTTACTTC 1560

GCGATCCTGC GCGCGGGTGC CGTCGGCGTG CCGCTCAACC CCGGTTCGGC GACGGCCGAG 1620

CTCGAGCACC CGCTGACCGA CAGCGGCGCC ACGGTGGTCG TCACCGACGC CGCCCAGGCG 1680

GCCCGGCTCC GGCTCGCGCC GCACGTCGAG CTGCTGGTGA CCGGCGACGA CGTCCCGGAG 1740

GGCGCCCACT CCTACGACGA ACTCGCCCTC AGCGAACCGG CCGAGCCCGC CGCGGACGAC 1800

CTCGAGCTCG ACGAGCCGGC GTGGATGTTC TACACGTCGG GCACGACCGG GCGCCCAAG 1860

GGCGTCGTGT CCACGCAGCG CAACTGCCTC TGGTCCGTCG CTTCCTGCTA CGTGCCGTTC 1920

CCCGGGTTGT CGGACCAGGA CCGGGTGCTC TGGCCGCTCC CGCTGTTCCA CAGCCTTTCG 1980

CACATCGCCT GCGTCCTGTC CGCCACCGTG GTCGGGGCCA GCGTCCGGAT CGCCGACGGC 2040

AGCTCCGCCG ACGACGTGAT GCGGCTGATC GAGGCGGAGA GCTCGACCTT CCTGGCCGGC 2100

GTGCCGACCA CCTACCACCA CCTGGTGCGG GCCGCCCGGC AGCGCGGTTT CTCCGCGCCG 2160

AGCCTGCGGA TCGGCCTGGC CGGGGCCGCG GTCCTCGGCG CCGGGCTGCG AAGCGAGTTC 2220

GAAGAGACCT TCGGGGTCCC GCTGATCGAC GCCTACGGCA GCACCGAGAC CTGCGGGGCG 2280

ATCACCATGA ACCCGCCGGA CGGCGCCCGC GTCGAGGGCT CGTGCGGCTT GGCCGTGCCG 2340

GGCGTCGACG TGCGGGTCGT CGACCCCGAC ACCGGGCTCG ACGTCCCCGC CGGCGAGGAG 2400

GGCGAGGTCT GGGTCAGCGG GCCGAACGTC ATGCTCGGCT ACCACAACAG CCCGGAGGCG 2460

ACCGCCGCGG CGATGCGGGA CGGCTGGTTC CGGACCGGGG ACCTGGCCCG CCGCGACGAC 2520

GCCGGTTACT TCACCATCTG CGGCCGGATC AAGGAACTCA TCATCCGCGG CGGCGCCGAAC 2580

ATCCACCCCG GCGAGGTCGA GGCGGTCCTG CGCACGGTCG ACGGCGTCGC GGACGCGGCG 2640

GTCGGCGGTG TGCCGCACGA CACGCTCGGC GAGGTGCCGG TCGCCTACGT CATCCCCGGA 2700

CCGACCGGTT TCGATCCTGC GGCGTTGATC GAGAAGTGCC GCGAACAGCT GTCCGCCTAC 2760

AAGGTGCCGG ACCGGATCCT CGAGGTCGCC CACATTCCCC GGACCGCGTC GGGCAAGATC 2820

CGGCGCGGGC TGCTGACCGA CGAGCCCGCG CAGCTGCGGT ACGCCGCGAC CGAACACGAG 2880

GAACAGTCCC GGCACGCCGA CGAGTCCGTC GCGGCGGCGC TGCGCGCGCG ACTGTCCGGT 2940

TTGGACGAAC GCGCCCAGTG CGAGCTCCTG GAAGACCTCG TCCGCACCCA GGCGGCCGAC 3000

GTGCTGGGGC AGCCGGTCCC GGACGGGCGT GCGTTCCGCG ACCTCGGCTT CACGTCGCTG 3060

GCCATCGTGG AGCTGCGCAA CCGGCTGACC GAGCACACCG GGCTCTGGCT GCCCGCCAGC 3120

GCCGTCTTCG ACCACCCCAC GCCGGCGGCG CTGGCCGCCC GCGTCCGGGC TGAGCTCCTC 3180

GGGATCACGC AGGCCGTCGC GGAGCCGGTC GTCGCGGCCG ACCCGGGCGA GCCGATCGCG 3240

ATCGTGGGGA TGGCCTGCCG CCTGCCGGGT GGCGTGGCGT CCCCGGAAGA CCTGTGGCGG 3300

CTGGTGGCCG AGCTCGTCGA CGCCGTTTCG GAGTTCCCCG GCGACCGCGG CTGGGACCTG 3360

GACAGCCTGA TCGACCCGGA CCGGGAGCGC GCCGGGACGT CGTACGTCGG CCAGGGCGGA 3420

TTCCTGCACG ACGCCGGCGA GTTCGACGCC GGGTTCTTCG GGATCTCGCC GCGTGAGGCC 3480

GTCGCGATGG ACCCGCAGCA GCGTTTGCTG CTGGAGACGT CGTGGGAGGC CCTCGAAAAC 3540

GCCGGAGTCG ACCCGATCGC GTTGAAGGGC ACCGACACCG GCGTGTTCTC CGGCCTCATG 3600

GACCAGGGGT ACGGGTCCGG CGCGGTGGCG CCGGAGCTCG AAGGTTTCGT CACCACCGGG 3660

GTCGCGTCGA GCGTGGCCTC GGGCCGGGTG TCGTACGTGC TGGGACTGGA AGGCCCGGCG 3720

GTCACCGTGG ACACCGCGTG TTCGTCGTCG CTGGTCGCGA TGCACCTGGC CGCGCAGGCC 3780

CTGCGGCAGG GCGAATGCTC GATGGCGCTC GCCGGCGGGG TCACGGTGAT GGCCACGCCG 3840

GGCTCGTTCG TCGAGTTCTC CGCCCAGCGG GCCCTGGCGC CGGACGGGCG CTGCAAGGCC 3900

TTCGCGGCGG CGGCCGACGG GACCGGCTGG TCCGAGGGTG TCGGCGTGGT CGTCCTCGAG 3960

CGGCTGTCCG TGGCGCGCGA GCGGGGCCAC CGGATCCTGG CCGTTTTGCG TGGCAGCGCG 4020

GTCAACCAGG ACGGCGCGTC CAACGGGCTC ACCGCGCCGA ACGGCCTCTC GCAGCAGCGG 4080

GTCATCCGCC GCGCGCTGGC CGCGGCCGGG CTGGCACCGT CCGATGTGGA CGTCGTCGAG 4140

GCGCACGGCA CCGGGACCAC GCTGGGTGAC CCGATCGAGG CGCAGGCCCT GCTGGCGACC 4200

TACGGCCAGG AGCGGAAGCA GCCGTTGTGG CTCGGTTCGC TCAAGTCGAA CATCGGCCAC 4260

GCGCAGGCGG CCGCGGGCGT TGCGGGCGTC ATCAAGATGG TGCAGGCGCT GCGGCACGAG 4320

ACCTTGCCGC CGACGCTGCA TGTCGACAAG CCGACTCTTG AGGTGGACTG GTCCGCCGGT 4380

GCCATTGAAC TGCTGACGGA GGCCCGTCCG TGGCCGCGCA ACGGCCGTCC GCGCCGGCC 4440

GGGGTGTCGT CGTTCGGCGT CAGCGGGACC AACGCGCACC TGATCCTGGA GGAGGCGCCG 4500

GCCGAGGAGC CGGTCGCTGC CCCGGAACTG CCGGTGGTGC CCCTGGTGGT GTCGGCGCGG 4560

AGCACGGAGT CGCTGTCCGG GCAGGCCGAG CGGCTGGCGT CCCTCCTGGA AGGGGACGTC 4620

TCGCTGACCG AGGTGGCCGG GGCGCTGGTG TCCCGCCGGG CGGTGCTGGA CGAGCGGGCC 4680

GTCGTCGTGG CCGGTTCGCG CGAGGAAGCC GTGACCGGGC TGCGGGCGCT GAACACGGCC 4740

GGTTCGGGGA CGCCGGCAA GGTCGTGTGG GTGTTCCCGG GGCAGGGGAC GCAGTGGGCC 4800

GGGATGGGCC GTGAGCTGCT GGCCGAGTCC CCGGTGTTCG CCGAGCCGAT CGCCGAGTGC 4860

GCGGCCGCGT TGGCGCCGTG GATCGACTGG TCGCTCGTCG ACGTCCTGCG CGGCGAGGGC 4920

GACCTGGGTC GGGTCGATGT GCTGCAGCCG GCCTGTTTCG CGGTGATGGT CGGGCTGGCT 4980

GCCGTCTGGG AGTCCGTGGG GGTCCGGCCG GACGCCGTCG TCGGGCACTC GCAGGGTGAG 5040

ATCGCGGCTG CCTGCGTTTC GGGGGCGTTG TCCCTCGAGG ACGCGGCGAA GGTCGTGGCC 5100

CTGGGCAGCC AGGCCATCGC GGCGGAACTG TCCGGCCGCG GCGGGATGGC GTCGGTCGCC 5160

CTGGGCGAGG ACGACGTCGT TTCGCGGCTG GTGGACGGGG TCGAGGTCGC CGCCGTCAAC 5220

GGCCCGTCGT CGGTGGTGAT CGCCGGGGAT GCCCATGCCC TCGACGCGAC CCTGGAAATC 5280

TTGTCCGGGG AAGGCATCCG GGTTCGGCCG GTGGCGCTGG ACTACGCCTC GCACACCCGG 5340

CATGTCGAGG ACATCCGCGA CACTCTTGCC GAAACCTTCG CCGGGATCAG TGCGCAGGCG 5400

CCGGCTGTGC CGTTCTACTC CACCGTCACG AGCGAGTGGG TGCCGGACGC GGGGGTGCTG 5460

GACGGCGCCT ACTGGTACCG GAACCTGCGC AACCAGGTCC GGTTCGGAGC GGCCGCGACG 5520

GCCCTGCTCG AGCAGGGCCA CACGGTGTTC GTCGAGGTCA GTGCGCACCC GGTGACGGTC 5580

CAGCCCTTGA GCGAGCTCAC CGGGGACGCG ATCGGGACAT TGCGGCGTGA AGACGGTGGC 5640

CTGCGGCGGT TGCTGGCTTC GATGGGTGAG CTGTTCGTCC GCGGCATCGA CGTGGACTGG 5700

ACGGCGATGG TGCCCGCGGC CGGCTGGGTC GACTTGCCGA CCTACGCGTT CGAACACCGG 5760

CACTACTGGC TCGAGCCCGC CGAGCCCGCT TCGGCCGGAG ACCCGCTGCT GGGCACAGTC 5820

GTCAGCACTC CCGGTTCGGA CCGACTCACC GCCGTGGCGC AGTGGTCGCG CCGGGCGCAG 5880

CCCTGGGCGG TGGACGGCCT GGTGCCGAAC GCGGCCCTGG TCGAGGCGGC CATCCGGCTC 5940